DISSERTATIONES SCHOLAE DOCTORALIS AD SANITATEM INVESTIGANDAM
UNIVERSITATIS HELSINKIENSIS

5/2019

LUCA TROTTA

Genetics of Primary Immunodeficiency in Finland

INSTITUTE FOR MOLECULAR MEDICINE FINLAND FIMM
FACULTY OF BIOLOGICAL AND ENVIRONMENTAL SCIENCES
DOCTORAL PROGRAM IN INTEGRATIVE LIFE SCIENCE
UNIVERSITY OF HELSINKI

GENETICS OF PRIMARY IMMUNODEFICIENCY IN FINLAND

Luca Trotta

Institute for Molecular Medicine Finland (FIMM)
University of Helsinki, Helsinki, Finland
Doctoral program in Integrative Life Science
University of Helsinki, Helsinki, Finland
Faculty of Biological and Environmental Sciences
University of Helsinki, Helsinki, Finland

ACADEMIC DISSERTATION

To be presented, with the permission of Faculty of Biological and Environmental Sciences, University of Helsinki, for public examination at Päärakennus Helsinki, Auditorium XII, Fabianinkatu 33, Helsinki, on 24 January 2019 at 12 noon.

Helsinki 2019

Supervised by:
Professor Janna Saarela, MD, PhD
Institute for Molecular Medicine Finland FIMM, University of Helsinki, Helsinki, Finland

Second supervisor:
Docent Mikko Seppänen, MD, PhD
Adult Immunodeficiency Unit, Department of Infectious Diseases, Inflammation Center, University of Helsinki and Helsinki University Hospital, Helsinki, Finland
Rare Disease Center, Children's Hospital, University of Helsinki and Helsinki University Hospital, Helsinki, Finland

Thesis advisory committee:
Docent Samuel Myllykängäs, PhD
Blueprint Genetics, Helsinki, Finland.
Institute of Biomedicine, University of Helsinki, Helsinki, Finland.
and
Docent Päivi Saavalainen, PhD
Research Programs Unit, Immunobiology, University of Helsinki, Helsinki, Finland.
Department of Medical and Clinical Genetics, University of Helsinki, Helsinki, Finland.

Reviewed by:
Professor Mika Rämet, MD, PhD
University of Oulu, Finland
University of Tampere, Finland
and
Adjunct Professor Hannu Turpeinen, PhD,
Senior Geneticist, Blueprint Genetics

Custos:
Professor Minna Nyström, PhD
Faculty of Biological and Environmental Sciences, Molecular and Integrative Biosciences Research Programme, University of Helsinki, Helsinki, Finland

Opponent:
Docent Asbjorg Stray-Pedersen, MD PhD
Norwegian National Unit for Newborn Screening, Division of Pediatric and Adolescent Medicine, Oslo University Hospital - Rikshospitalet, Oslo, Norway.

ISBN 978-951-51-4783-7 (Paperback)
ISBN 978-951-51-4784-4 (PDF)
ISSN 2342-3161 (Print)
ISSN 2342-317X (Online)

Helsinki 2019

TABLE OF CONTENTS

LIST OF ORIGINAL PUBLICATIONS

The presented thesis is based on the following publications.

(I) Trotta L, Hautala T, Hämäläinen S, et al. Enrichment of rare variants in population isolates: single *AICDA* mutation responsible for hyper-IgM syndrome type 2 in Finland. *European Journal of Human Genetics*. 2016;24(10):1473-1478. doi:10.1038/ejhg.2016.37.

(II) Trotta L, Martelius T, Siitonen T, Hautala T, Hämäläinen S, Juntti H, et al. ADA2 deficiency: clonal lymphoproliferation in a subset of patients. *J Allergy Clin Immunol* 2018 Jan 29. doi: 10.1016/j.jaci.2018.01.012.

(III) Trotta L, Norberg A, Taskinen M, et al. Diagnostics of rare disorders: whole-exome sequencing deciphering locus heterogeneity in telomere biology disorders. *Orphanet Journal of Rare Diseases*. 2018;13:139. doi:10.1186/s13023-0

AUTHOR'S CONTRIBUTION

I-III: Participated in the conception and design of the works in collaboration with the co-authors. Established the analytical workflow for the whole-exome sequencing data. Generated the annotations of the variant data in collaboration with the co-authors. Analysed the genetic data. Provided the interpretation of the clinical significance of the identified variants. Validated the genetic findings. Executed the population-level analysis. Drafted the manuscripts, revised and finalised them according to the feedback from the co-authors.

ABBREVIATIONS

ACD	acd shelterin complex subunit and telomerase recruitment factor
A	adenosine
ACMG	American College of Medical Genetics and Genomics
ADA2	adenosine deaminase 2
AICDA	activation induced cytidine deaminase
bp	
C	cytosine
CD4	cluster of differentiation 4
cDNA	complementary DNA
CECR1	cat eye syndrome chromosome region, candidate 1
CMV	cytomegalovirus
CNV	copy number variant
CTC1	cst telomere replication complex component 1
DADA2	deficiency of ADA2
DKC	dyskeratosis congenita
DKC1	dyskerin pseudouridine synthase 1
DNA	deoxyribonucleic acid
G	guanine
GWAS	genome-wide association study
HHS	Hoyeraal-Hreidarsson syndrome
HIGM2	hyper-IgM syndrome type 2
HSCT	hematopoietic stem cell transplantation
IgA	immunoglobulin A
IgE	immunoglobulin E
IgG	immunoglobulin G
IgG2	immunoglobulin G class 2
IgM	immunoglobulin M
InDELS	insertions or deletions
MAF	minor allele frequency
Mb	megabases
MMR	measles, mumps and rubella vaccine
ng	nanograms
NGS	next-generation sequencing
NHP2	NHP2 ribonucleoprotein
NK	natural killer cells
NOP10	NOP10 ribonucleoprotein
PARN	poly(A)-specific ribonuclease
PCR	polymerase chain reaction
PIDs	primary immunodeficiency disorders
RNA	ribonucleic acid
RTEL1	regulator of telomere elongation helicase 1
RTL	relative telomere length
SCID	severe combined immunodeficiency
SD	standard deviation
SISu	Sequencing Initiative Suomi
T	thymine
TBDs	telomere biology disorders
TERC	telomerase RNA component
TERT	telomerase reverse transcriptase
Th17	T helper 17
TINF2	TERF1 interacting nuclear factor 2
WES	whole-exome sequencing
WGS	whole-genome sequencing
WRAP53	wd repeat-containing antisense to TP53

ABSTRACT

Primary Immunodeficiency (PIDs) categorize a broad and heterogeneous group of inborn immunity errors. Despite being generally quite rare, PIDs collectively account for consistent morbidity and mortality. Currently, more than 350 monogenic PIDs have been recognised to embody clinical phenotypes ranging from life-threatening infections to autoimmune/inflammatory diseases, allergies and/or malignancy. Many PIDs display genetic and allelic heterogeneity with an overlap of symptoms among different syndromes, often making diagnoses challenging.

In the past few years, advancements in genomic technologies have revolutionised the world of genetic testing, and currently, next-generation sequencing (NGS)-based approaches are widely applied to routine genetic diagnostics of human disorders. Among the different methods, whole-exome sequencing (WES) proved highly efficient in revealing the genetic variants behind rare disorders. Despite only targeting the coding region of the genome, WES covers the largest share of causal variants in severe monogenic diseases.

To further depict the genetics of PIDs, a WES-based approach was carried out, targeting the possible disease-causing variants in Finnish subjects lacking a clinical diagnosis. The cohort included patients with a clinical suspicion of immune or/and haematological disorders (n= 212). The wide spectrum of phenotypes included severe early-onset disorders (e.g. severe compound immunodeficiency, SCID), known PIDs (e.g. antibody deficiency or telomere biology disorders, TBDs) and clinically undiagnosed conditions.

In the first study, a Finnish founder mutation in the *AICDA* gene was identified in patients affected by hyper-IgM syndrome type 2 (HIGM2). The disease is a primary antibody deficiency characterised by early-onset recurrent infections, autoimmunity and an absence/low levels of IgG, IgA and IgE but elevated/normal levels of IgM. The retrieved ancestral founder allele is significantly enriched in Finns compared to other European populations (38.56-fold) and has accounted for all HIGM2 cases diagnosed in Finns thus far.

In the second study, biallelic *ADA2* mutations that cause a deficiency of adenosine deaminase 2 (DADA2) were identified in seven PID patients, all sharing one of the causal variants, which were significantly enriched in Finns (3.31-fold). DADA2 was originally associated with systemic autoinflammation, polyarteritis nodosa-type vasculitis and mild immunodeficiency. Only a fraction of the identified DADA2 patients presented with vasculopathies. In addition, recurrent haematological manifestations are noted, and for the first time, the occurrence of lymphoproliferation is described for some of the patients, expanding the phenotypic spectrum of DADA2.

Finally, novel causal variants in telomere biology disorders (TBDs)-associated genes (*DKC1, TERT* and *RTEL1*) were identified in three families with heterogeneous phenotypes that lacked the classic clinical pathognomonic signs of telomeropathies. The phenotypes ranged from mild signs of Dyskeratosis congenita (DKC) to SCID. The genetic diagnosis was confirmed by an assessment of shortened telomere lengths in patients. In addition, the spectrum of TBD-associated phenotypes was enlarged, showing variable degrees of cytopenia in some patients.

This work attests to the validity of clinical WES testing to identify rare disease-causing variants despite the heterogeneous and/or atypical clinical presentations of PIDs. The achievement of a genetic diagnosis allowed for updating the spectrum of reported phenotypes as well as including atypical clinical presentations that might have otherwise remained undiagnosed. In addition, the enrichment of some rare PID-causing mutations in Finland has been depicted, highlighting the correlation of the population history with the distribution of rare deleterious variants of clinical relevance.

1 INTRODUCTION

In the past two decades, technological breakthroughs in DNA sequencing technologies have advanced human genome deciphering, and large-scale genetic diagnostic tests have been developed. The current next-generation sequencing (NGS) methodologies have substantially increased the sequencing yield and have lowered the cost per base (Buermans et *al.* 2014). Consequently, the possibility to employ high-throughput and cost-efficient DNA sequencing strategies has dramatically transformed the diagnostic approach to genetic diseases (Seleman et *al.* 2017). Currently, the routine use of NGS-based methods has quickly spread in clinical practices, setting the stage for the development of personalised medicine approaches (Rabbani et *al.* 2014).

The identification of causal variants in inherited diseases is the outset for either preventive and treatment strategies or for terminating the diagnostic odyssey that many patients endure (Jamuar et *al.* 2015). The application of NGS methods in clinical settings varies for different diseases. In the field of Mendelian disorders, the increasing use of whole-exome sequencing (WES) and whole-genome sequencing (WGS) has largely driven the identification of underlying biological mechanisms. In particular, providing a suitable balance between costs and sequencing spans, WES has emerged as a highly effective strategy for deciphering the genetics of rare diseases (Shen et *al.* 2015).

Rare conditions collectively exert a large burden on society (Dodge et *al.* 2011; Chong et *al.* 2015, von der Lippe et *al.* 2017). In addition to a large number of affected individuals, diagnostic and therapeutic limitations often contribute to a negative impact on patients' lives and on public health resources. The low prevalence of many Mendelian conditions reduces awareness and knowledge among healthcare professionals as well as society (von der Lippe et *al.* 2017). Moreover, the clinical description of novel syndromes could be biased by their low prevalence, potentially reducing the coverage of the phenotypic spectrum and the potential to identify individuals with overlapping features. In addition to prevalence-related issues, Mendelian disorders are often characterised by genetic and allelic heterogeneity and an overlap of symptoms among different conditions, which can make diagnostic outcomes challenging. Additional challenges emerge due to population migrations because conditions once considered rare or linked only to specific ethnic groups have been identified in clinics (Dodge et *al.* 2011).

Primary immunodeficiencies (PIDs) comprise the heterogeneous spectrum of alterations of the development and/or functions of the immune system. The degree of severity of the disorders varies considerably, but PIDs collectively cause consistent morbidity and mortality. The burden of PIDs on the society is further increased by the diagnostic hurdles that typically result from tackling heterogeneous clinical phenotypes (Bousfiha et *al.* 2013, Seleman et *al.* 2017). In the era of NGS, the number of described PIDs has increased, and there are current descriptions of over 350 individual conditions linked to alterations in more than 350 genes (Picard et *al.* 2018). Nonetheless, not all the molecular defects that cause PIDs have been identified, and the phenotypic spectra of numerous syndromes remain only partially known. Recent estimations suggest that PIDs are even

more prevalent than assumed, highlighting the need for the improvement of management strategies provided by public health systems (Bousfiha et *al.* 2013).

The aim of the study was to elucidate the genetic bases of immune and/or haematological conditions overlapping PIDs in Finnish individuals who lacked a conclusive diagnosis. Therefore, WES-based methods were applied for the routine diagnostics of potential PIDs. The identification of PID-causing variants could reduce challenges typically related to the diagnostic procedures of several rare syndromes. Moreover, implementing the genotype-phenotype correlation in rare PID entities and determining the distribution of rare detrimental variants in Finland can allow for tailoring diagnostic and treatment strategies, even from a population perspective.

2 REVIEW OF LITERATURE

2.1 The human genome

2.1.1 Structure

The human genome is comprised of nuclear and mitochondrial deoxyribonucleic acid (DNA) sequences. Respectively formed by over three billion DNA base-pairs (bp) contained in the cellular nuclei and 16,6 DNA kilobases (kb) within the mitochondria, the overall DNA sequence carries intrinsic information of human biology.

At the molecular level, DNA is composed of two strands coiled around each other in the shape of a double helix. Each strand consists of multiple basic units defined as nucleotides, which consist of a deoxyribose-phosphate backbone and one of the four nitrogenous bases. When forming a double-helix, the bases on adjacent poly-nucleotide chains pair reciprocally (adenosine [A] with thymine [T] and cytosine [C] with guanine [G]).

DNA chains are organised into complexes with proteins defined histones. The resulting chromatin is tightly packaged in increasing levels of complexity, from nucleosomes (DNA segments of ~150bp are wrapped around eight histones) to highly-coiled sub-cellular chromosomal structures (Figure 1). The nuclear genome is organised into 23 pairs of chromosomes (22 autosomes and one pair of sex chromosomes, X and Y) due to its diploid nature because a copy of the genome (haploid) is inherited from each parent. The mitochondrial genome is more simply structured with one single circular chromosome present in multiple copies in each mitochondrion.

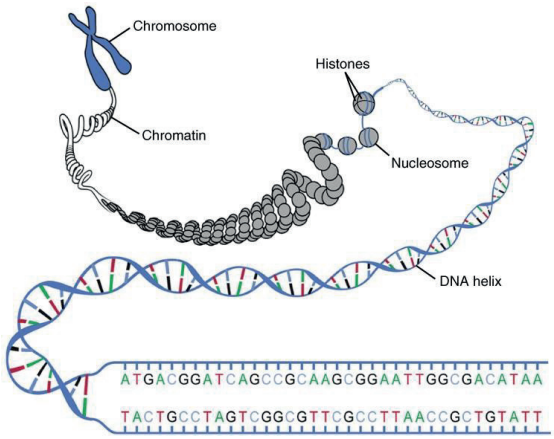

Figure 1. DNA structure and organization: from double-helix to chromosomes (from Wikimedia Commons, the free media repository).

Functionally, DNA sequence elements are organised into segments displaying a specific biochemical function, also defined as coding regions (protein or non-coding RNA) (ENCODE Project Consortium, 2012). Among functional genomics, unit protein-coding genes have been the most extensively studied and are estimated to total 20.000-25.000 (Ezkurdia et *al.* 2014). The expression of protein-coding genes is controlled in eukaryotic organisms by complex regulatory mechanisms layered on hierarchical levels, ascribable to the transcription of individual coding units, the structural organisation of chromatin and the epigenetic regulation of gene expression (van Driel et *al.* 2003).

The completion of the Human Genome Project over a decade ago represented a turning point for genetic and genomic studies. The resulting first complete draft of the reference genome sequence allowed for establishing baselines for comparisons with the sequence data generated by following individual re-sequencing studies (Lander et *al.* 2001, International Human Genome Sequencing Consortium, 2004). Afterward, a succession of large-scale genotyping and sequencing efforts produced data included in publicly available registers of population-level human variations (e.g. The 1000 Genomes Project Consortium, 2015; Exome Variant Server: the Exome Aggregation Consortium, ExAC, 2016; the Genome Aggregation Database 2016, gnomAD, Lek et *al.* 2016). This increasing volume of data has supported population genetics studies, and it contributes to the continuous process of deciphering the features of genetic variations, especially related to human diseases.

2.1.2 Variation

All humans share a full sequence identity for about 99,9% of their genomes (Venter et *al.* 2001). The remaining 0,1% of the sequence variation could contribute to shaping the differences among individuals along with environmental factors. Any difference in sequence among the (consensus) reference genome and the nucleotide sequences generated by subsequent genomic studies is considered as genetic variant. A genetic variation is determined by permanent changes in the DNA sequence of a genome. It can occur in different forms and can involve different regions of the genome with different extent, if any, on human phenotypes. In addition to homologous genetic recombination (occurring during meiosis), sequence changes arise in the genome through mutation processes. Mutations can be spontaneous, such as through ensuing errors in DNA replication or repair processes or can be induced by environmental factors (e.g. chemicals, radiations).

Overall, mutation frequency is not uniform along the genome due to mutational properties and/or selective pressures (Hastings et *al.*1991; Nachman et *al.* 2000; Tiemann-Boege et *al.* 2002; Goriely et *al.* 2003; Choi et *al.* 2008; Qin et *al.* 2007; Arnheim et *al.* 2009). Certain nucleotide positions, defined as mutation 'hotspots', were shown to be more prone to variations, and a higher rate of variation often reflects specific nucleotide-sequence features, such as repetitive sequences (e.g. homonucleotide runs, microsatellites) or CG-dinucleotide rich regions (CpG sites) (Rogozin et *al.* 2003; Kong et *al.* 2012).

Genetic variation can be viewed from the nucleotide-level or the chromosome-level perspective. The changes at the nucleotide level, namely the monomers of DNA strands, are

defined as single-nucleotide variants (SNVs). Small insertions or deletions (indels) represent gains or losses of genetic sequences usually smaller than 50bp, while structural variations (SVs) include larger differences (even ≥1kb). Copy-number variants (CNVs) are duplications or deletions of DNA segments, which are present in a number of copies different than the reference. Other macroscopic forms of structural variations can affect the number and/or the genetic content of one or more chromosomes. Based on the number, SNVs represent the largest category of variations, taking place on average each 8bp within the exonic regions according to updated estimates (The 1000 Genomes Project Consortium 2015; Lek et *al.* 2016).

From a technical perspective, the detection of some forms of variations remains problematic, including small CNVs (<500bp), the insertion of sequences different from the reference, subtle variations in the number of copies, highly-represented dispersed repeats (e.g. *Alu* elements, LINEs) and CNVs located on the Y chromosome or heterochromatic regions (Wheeler et *al.* 2008; Conrad et *al.* 2010).

2.2 Genetic variations in human diseases

Any individual human genome presents an estimated amount of ~4,1-5 million variation sites in comparison to the reference. This variation load is mostly (>99,9%) represented by SNVs and short indels (Figure 2) (The 1000 Genomes Project Consortium 2015). Overall, more than 100 million variants have been listed in the dbSNP database of human sequence variations (Sherry et *al.* 2001; The 1000 Genomes Project Consortium 2015). Recent large-scale sequencing projects provided a broader view of existing genomic variability, showing a significant quote to be tolerated or not linked at any rate to (manifest) deleterious effects (The 1000 Genomes Project Consortium 2015; Lek et *al.* 2016). A fraction of the depicted genetic variations could result from positive selection due to beneficial evolutionary effects, but not all genetic variants are neutral or are not linked to any phenotypic effect.

Currently, according to the Online Mendelian Inheritance in the Man OMIM® database (https://omim.org/; Hamosh et *al.* 2005), variants in 3874 genes are linked to more than 6000 clinical phenotypes, including both single-gene Mendelian disorders and susceptibilities to cancer and complex diseases. Focusing on the disease-associated variants, ~180.000 entries are listed in the Human Gene Mutation Database (HGMD®) (Krawczak et *al.* 1997).

The effect of mutations occurring in a genome can be evaluated at the evolutionary level as neutral, beneficial or deleterious based on the effects of the nucleotide changes on the resultant gene products. If the function of the gene product is abolished, the causing-variants are defined as loss-of-function (LoF) or hypomorphic if the repercussion is only partial. A dominant-negative effect occurs when the mutated product can affect the wild-type product at the functional level. Other variants can result in aberrant or heightened activity, causing the so-called gain-of-function (GoF) effect.

If the variants occur in protein-coding genes, their consequences can be evaluated on the resulting protein structure. Variants that do not alter the amino acid sequence are defined as silent (or synonymous). If the result is instead the substitution of one amino acid with a different one,

the variants are defined as missense (or non-synonymous). Different type of variants which could result in reduced or abolished protein function, defined as nonsense (introduction of a premature stop codon), splice site (removing physiological splice sites or inserting aberrant ones) or frameshift indels (interruption of the amino acid sequence reading frame) can be grouped in the category of LoF. Indels that do not induce frameshifts (e.g. inserting or deleting triplets of nucleotides) are defined as in-frame. In addition to the inherited variation, *de novo* mutations can play a significant role in disease pathogenesis. *De novo* mutations manifest without apparent parental inheritance, either arising in the germ cell of one of the parents or in the fertilised egg during early embryogenesis, with an overall estimated frequency of $\sim 1{,}5\times10^{-8}$ SNVs per site (Turner et *al.* 2017). When *de novo* mutations occur after conception in cell lines other than the germline, they are defined as somatic, and only that specific cell population will carry it. Somatic mutations contribute to several types of cancer disorders.

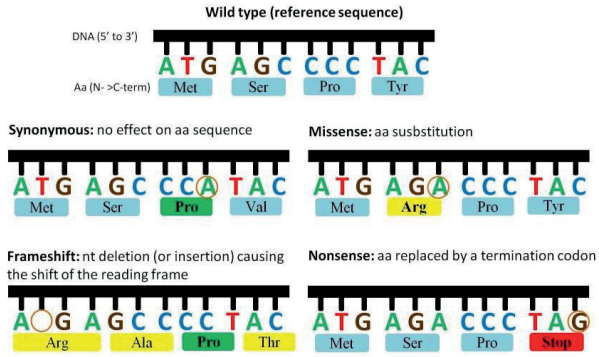

Figure 2. Consequences of point mutations (SNVs) on the protein.

2.3 Genetics of rare disorders

Rare genetic diseases are medical conditions that affect a small fraction of the population. According to the European classification, a disease is labelled rare if it affects less than one individual of 2000 (http://www.orpha.net/consor/cgi-bin/index.php, May 2017), while in the US, it is rare if the total of affected patients is less than 200.000 (http://www.fda.gov, May 2017) (Pogue et *al.* 2017). Despite being in general quite rare, Mendelian diseases collectively exert a large burden on public health due to the large number of patients affected. A total of 6000-8000 mostly Mendelian (or monogenic) rare diseases have been estimated. Altogether, over 30 million people in Europe and 25 million in the US suffer from severe and/or chronic manifestations of Mendelian conditions, often with absent or limited treatment options (Haffner et *al.* 2002; Dodge et *al.* 2011; von der Lippe et *al.* 2017).

The limited number of described clinical cases reduces knowledge both at the professional and at the society level, resulting in long, erroneous and/or inconclusive diagnostic procedures. Moreover, the more severe phenotypes are detected first, usually having a greater contribution to the description of disease-associated phenotypes and potentially hindering future correlations with milder traits. In addition, the diagnostic process could be deceptive in the presence of novel genetic causes that are not targeted by the genetic tests. Along with low prevalence-related issues, the overlap of clinical symptoms among different conditions and the high genetic heterogeneity reported can further complicate the diagnosis of rare diseases.

From a genetic perspective, rare disorders are mostly caused by fully-penetrant single-gene alterations, with (coding-) mutations in over 3000 genes previously disclosed (Table 1, McKusick et *al.* 2007). According to the Mendelian segregation, the phenotypes can be inherited with autosomal-recessive, autosomal-dominant or X-linked recessive patterns. Alternatively, in case of incomplete penetrance the disease-causing variants can be inherited from apparently unaffected parents. Recessive disorders are mostly caused by LoF or hypomorphic variants, while dominant disorders are mostly due to GoF and dominant-negative variants. X-linked recessive disorders manifest only in male individuals harbouring one copy of the causative variant, although in some cases mild symptoms can be observed in female carriers.

Recent population-level sequencing studies showed an uneven load of potentially detrimental variants (LoF, missense) across distinct genes, suggesting that different degrees of selection can determine the shift from the estimated mutation rate (Lek et *al.* 2016). Accordingly, disease-causing genes have been shown as less tolerant to functional genetic variations than others not associated with any known disease (Petrowski et *al.* 2013).

Phenotypes with known molecular basis*	6,153
Genes with phenotype-causing mutation	3874
Genes linked to 1 phenotype	2,652
Genes linked to 2 phenotypes	729
Genes linked to 3 phenotypes	259
Genes linked to 4 or more phenotypes	234

Table 1. Gene statistics from the Online Mendelian Inheritance in the Man OMIM® database (https://omim.org/) (updated January 31st, 2018).
*Phenotypes including single-gene Mendelian disorders and traits, susceptibilities to cancer and complex disease variations that lead to abnormal but benign laboratory test values ('non-diseases') and blood groups.

2.4 Genetic testing in Mendelian diseases

Given their mostly monogenic aetiology, Mendelian disorders have always represented a forefront for genetic association studies, and thus far, 3874 genes have been linked to rare monogenic

syndromes (http://www.orpha.net/consor/cgi-bin/index.php, May 2017). Understanding the genetic bases of rare diseases can improve the clinical practices, and the genes-phenotypes correlation can shed light on the biological mechanisms underlying clinical conditions (Figure 3; Dodge et *al.* 2011). Nonetheless, the low prevalence of Mendelian disorders has always represented an obstacle to identifying the causes. Another limitation is genetic heterogeneity, which has recently been shown to reduce the diagnostic rates (Chong et *al.* 2015). Deciphering the molecular bases of Mendelian disorders can enable discerning distinct conditions not previously recognised as single entities and that have been included in more complex phenotypes, thus reducing the difficulties that patients experience as well as difficulties related to medical practices (Dodge et *al.* 2011; von der Lippe et *al.* 2017).

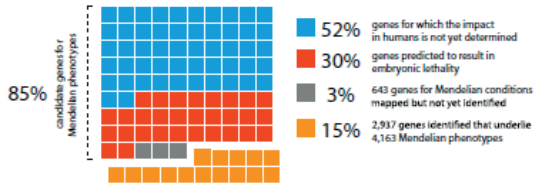

Figure 3. The link between protein-coding genes (~19,000) and Mendelian phenotypes (~3,000). Adapted from Chong et *al.* (2015).

Beginning in the 1980s, the discovery of genes associated with Mendelian disorders resulted from indirect positional cloning and genome-wide linkage analyses or candidate-gene approaches (Figure 4, Chong et *al.* 2015). If candidate-gene approaches clearly rely on previous biological assumptions, the main constraint of positional cloning occurs due to the limitation of familial traits with many affected members, therefore neglecting rare sporadic cases. In addition, the genetic heterogeneity of Mendelian diseases is a disadvantage for linkage studies, which have typically been more successful in populations presenting a reduced load of genetic variations, such as the Finnish isolate (Peltonen et *al.* 1998). After indirect methods were used for two decades, Sanger sequencing represented the standard method for Mendelian disease testing, allowing for the analysis of single genes (or sections of them) with high accuracy (Sanger et *al.* 1975). However, this approach is expensive and laborious if applied to multiple or larger targets, which is required for genetic heterogeneous disorders.

Technological breakthroughs over the past decade have resulted in drastic improvements of genetic testing methods, increasing the content and automation level of sequencing processes while reducing the cost per base. This progress is generically attributed to next-generation sequencing (NGS). Introduced in 2005, NGS-based methods have revolutionised the field of gene discovery, allowing for parallel sequencing of multiple genes, the whole protein-coding region (whole-exome sequencing, WES) or genomes (whole-genome sequencing, WGS) (Jamuar et *al.* 2015). The ever-increasing use of WES and WGS has dramatically advanced the identification of the genetic causes of Mendelian phenotypes (Chong et *al.* 2015). Given the impact of DNA sequencing, the development of personalised medicinal approaches is expected in upcoming years (Rabbani et *al.*

2014). Nonetheless, the use of NGS is still highly demanding in terms of storing, processing, analysing and interpreting the massive amount of continuously generated genomics data.

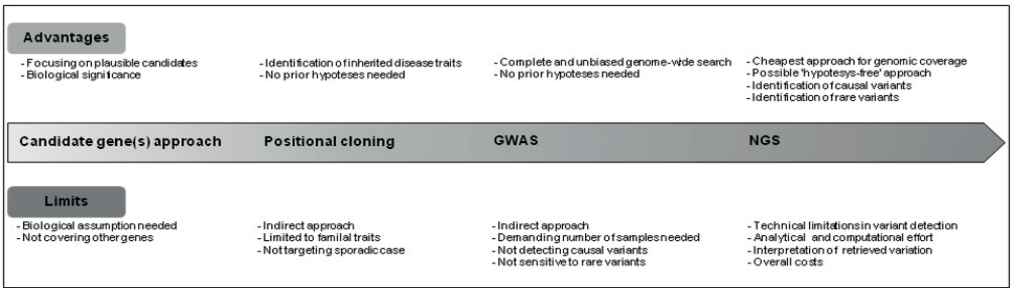

Figure 4. Main advantages and limitations of the most commonly used methods (from the 1980s to the present) for the discovery of genes associated with Mendelian disorders.

2.5 Next-generation sequencing applications in the study and diagnostics of rare diseases

The identification of disease-causing variants provides support for diagnostic, preventive and therapeutic strategies. Consequently, the use of NGS-based methods has increased in clinical settings despite the associated drawbacks, namely high costs and analytical and ethical issues related to the generated genetic data (Jamuar et *al.* 2015) (Figure 5). A crucial aspect of the adoption of NGS is the selection of the different testing applications that are most relevant to the specific diagnostic requirements. Gene-panels (targeted sequencing), WGS and WES are among the most widely adopted methods in clinical genetics.

Targeted approaches are applied to clinical phenotypes suggestive of a specific genetic aetiology according to the so-called 'phenotype first' approach. The targeted genes (scaling up to hundreds) are probed using panels, generally screening only the coding sequence of the candidates. In specific cases, the coverage of the disease-target genes can be integrated using complementary technologies (i.e. Sanger sequencing, long range PCR) (Jamuar et *al.* 2015). Targeted approaches are time- and cost-effective due to the low yield of generated data. In addition, the coverage of the established disease-causing genes is higher than for WES (Oishi et *al.* 2014). On the other hand, main limitations of targeted approaches lie in the reliability of the (established) genotype-phenotype correlations and in the needs of expert and accurate phenotype recognition. Other downsides involve limitations in deciphering genetically heterogeneous conditions and in identifying new disease genes.

WGS creates the possibility to identify variants genome-wide without requiring target selection or enrichment strategies. Compared with other NGS-based methods, WGS provides more uniform coverage and produces the highest diagnostic yield including genomics or variations outside the reach of WES (intronic, noncoding RNA genes, small CNVs, exonic SNVs due to WES coverage and enrichment issues). However, the limitations still include the increased cost and the demanding

requirements for data analysis, interpretation and storage (Lionel et *al.* 2017). Because the cost of WGS still prevents its routine application in diagnostic setups, WES currently represents an effective method to investigate the coding regions of a genome, and it is widely adopted in research and diagnostic contexts (Rabbani et *al.* 2014). Targeting only about 1% of the human genome, WES has become the most cost-effective approach because the largest part of the currently identified Mendelian phenotypes has originated from coding-sequence alterations (Chong et *al.* 2015; Shen et *al.* 2015). In addition to the discovery of novel disease-causing genes, WES can successfully identify causative variants in heterogeneous disorders (Rabbani et *al.* 2014). Nonetheless, despite targeting whole protein-coding regions, comprehensive WES coverage estimates are between 85–95%, indicating a potential lack or incomplete coverage of phenotype-related genes (Jamuar et *al.* 2015). Moreover, technical limitations and/or specific genomic features (e.g. repetitive sequences, homopolymers, high-GC content) could hamper the uniformity and the average coverage of the targeted exome.

WES and WGS can be used as 'genotype first' approaches in which a diagnosis can be established based on the molecular findings without any previous hypothesis or acknowledgment of the biological implication(s) of the disease-causing gene(s). The 'hypothesis-free' approach allows for detecting disease-causing mutations in cases of clinical presentations that do not conclusively match with known conditions, and it also targets a wider set of low-prevalence conditions.

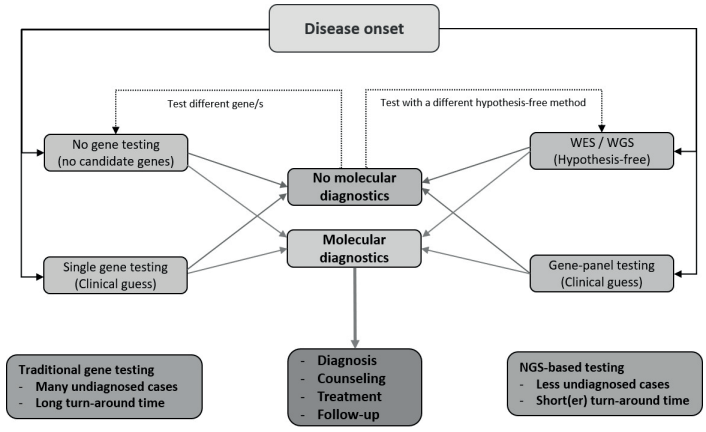

Figure 5. Genetic diagnostics flowchart before and at the time of next-generation sequencing.

2.6 Sequencing and processing of the obtained data

In the analytic phase of NGS, sequences are generated and analysed as depicted in Figure 6. The analysis of NGS data requires significant bioinformatics and computational processing, usually

enclosed in informatics pipelines designed and tailored based on their clinical applications. Currently, a large set of public or in-house software tools is available to process sequencing data, of which the Genome Analysis Toolkit (GATK) represents one of the most widely used in research (McKenna et *al.* 2010).

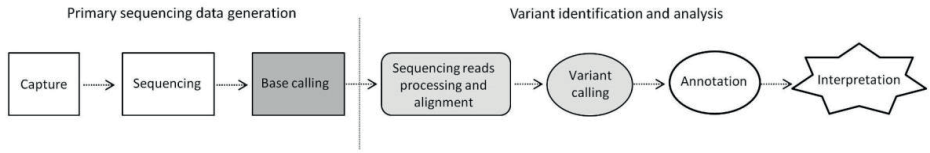

Figure 6. Flowchart of NGS-data generation and bioinformatics analysis.

2.6.1 Sequence data generation and processing

Prior to sequence generation, the template material is used to generate a 'library' of size-selected DNA fragments. For targeted or WES analyses, it is necessary to capture or enrich the region of interest (ROI) to enable the sequencing process. No enrichment is required for WGS. The use of library-specific oligonucleotide labels (tag or index) allows for labelling each fragment with the patient-specific specimen, permitting the analysis of multiple sample libraries simultaneously (multiplexing).

Following library generation, the clonal amplification of the DNA fragments and massive parallel sequencing are performed. The base-calling processes generate digital representations of short nucleotide sequences defined as 'reads', whose quality is controlled after the base-calling process according to the criteria established in each sequencing centre. In the following stages, the sequencing reads are mapped and aligned to the genome for the identification of sequence variants. The mapping quality of each read is assessed, and any read pair or single read mapped to multiple genomic positions is removed. Next, the refinement of the alignment of read mapping in the proximity of indels sites and the removal of any potential PCR-resulting bias (duplicate reads) are performed.

2.6.2 Detection of sequence variants

The major tasks involved in NGS are the accurate detection of SNVs and small indels located in the ROI, and distinguishing between true variants and sequencing or analytic errors. Therefore, after alignment, the coverage over the ROI and the quality of the whole process are assessed. To ensure high accuracy, a coverage threshold can be used to recalculate the resulting variants according to the quality values. Variant calling involves multi-step processes that can generally be performed by a variety of tools. A recent comparison of different methods suggested the joint use of both GATK and SAMtools to improve variant detection during a WES analysis of human diseases (Kim et *al.* 2017). GATK is an assortment of tools used for the analysis of genomic data, which was developed and released by the Broad Institute (McKenna et *al.* 2010). The GATK

variant calling process utilises the HaplotyperCaller tool (DePristo et *al.* 2011). SAMtools is used to process SAM and BAM files originated by alignment processes. It includes BCFtools, which contains utilities for variant calling processes (Li et *al.* 2009). Because the accurate calling of indels still presents challenges, additional programmes are available for detection. Dindel is a tool specifically developed by the Wellcome Trust Sanger Institute for the detection of indels (Albers et *al.* 2011).

2.6.3 Variant annotation

Variant calling processes result in a list of individual genotypes that differ from the reference genome. The amount of genetic information produced depends on the different sequencing platforms and pipelines used for data generation and analysis. The next stage involves the interpretation of the clinical significance of the detected variants. To manage the (usually) large list of variations generated, several computer-based annotation programmes allow for the prioritisation and filtering of potential causal variants. A common approach is linking the retrieved variants to information from genetic variation public databases (Pabinger et *al.* 2014).

To interpret causal variants, it is important to determine whether they are located in coding genes and whether they potentially affect the encoded products. Annotation programmes are usually used to mine gene- and population-level data as well as disease-specific information. Different predictions of the functional consequences caused by variants can result from different approaches, including sequence- and region-based analyses and the evaluation of the resulting structure of the affected protein. The predictions are used to classify the variants based on their supposed deleteriousness (DePristo et *al.* 2011). In addition, recent massive sequencing efforts released publicly available registers of population-level human variations (see paragraph 2.1.1), which allows for discarding a large part of the retrieved variants. Each application for variant annotation refers to one or multiple gene annotation sets (RefSeq and Ensembl among the most common). Several annotation tools can be used as applications without the need to install them on local machines (Pabinger et *al.* 2014). Among the many tools available, ANNOVAR is one of the most widely adopted tools for annotations and for predictions of the functional consequences of the identified variants (Wang et *al.* 2014; Chang et *al.* 2012).

2.7 Interpretation of sequencing results and identification of disease-causing variants

2.7.1 Filtering of whole-exome sequencing data

Regardless of the application used, NGS generates a number of variants which can exceed the analytic possibility of the users. Filtering processes are therefore applied to reduce the number of variants for further rounds of analysis. These processes are modulated based on the disease features

and the analytic plans of the executing laboratories (Figure 7). The first selection criterion for variant filtering is the effect on the resulting gene product. Non-coding (inter-genic or intronic) variants are typically discarded, and nonsense, missense, synonymous, splice site SNVs and indels are included instead. A large part of the remaining variants is pruned by filtering based on the population frequency. The variants that are common in the general population are usually discarded. The threshold for the allelic frequency is established according to the estimated inheritance pattern, the disease frequency and/or the expected disease allele frequency. Usually, the set is 0,01 for recessive diseases and lower for dominant ones. To date, the largest available register of population-level human variations is the Genome Aggregation Database (gnomAD, 2016, Lek et *al.* 2016).

The prioritisation process of causal variants continues based on reported links with possible phenotypic effects. For example, established sources of gene-disease information are OMIM® and the Human Gene Mutation Database (HGMD®) (Krawczak et *al.* 1997; Hamosh et *al.* 2005). Additional filtering can be performed based on the assumed inheritance pattern or phenotype segregation (Figure 8), and when available, the prediction of variants' pathogenicity.

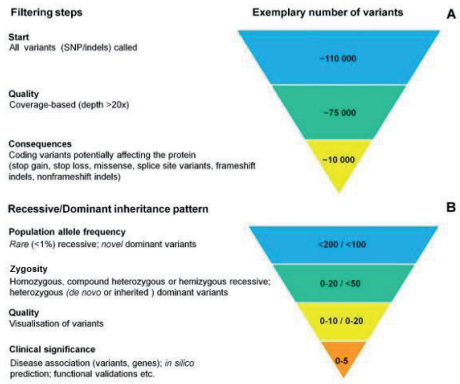

Figure 7. Filtering steps of whole-exome sequencing data to search for the disease-causing variant(s).
Panel A shows the filtering steps based on the quality and the consequences of the retrieved variants. Panel B lists the filtering applied in the function of assumed (or inferred) recessive or dominant (and *de novo*) inheritance patterns. The typical number of variants retrieved is based on the specific inheritance pattern.

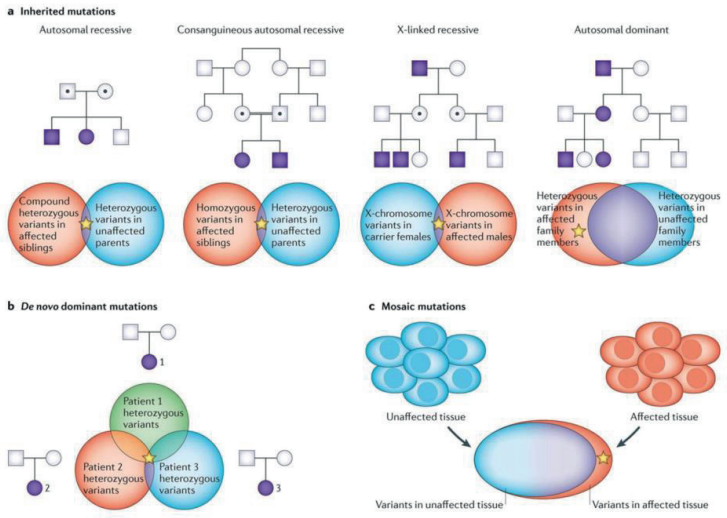

Figure 8. Whole-exome sequencing-based approach for gene identification in rare disorders. Adapted from Boycott et *al.* (2013).
The panels show pedigrees exemplary for different segregation patterns of rare diseases. Male individuals are indicated by squares and females by circles. Purple-coloured symbols specify affected individuals, while dots in the squares/circles indicate the carriers of a rare disease-causing mutation. Stars in the circles below the pedigrees indicate the range in which the possible disease-causing mutations are searched. The selection of the appropriate strategies for disease-gene identification is based on the expected mode of inheritance of the disease variants, whether the mutations are inherited, *de novo* or mosaic.
A. Research of inherited disease-causing variants. For autosomal recessive disorders, the analysis of affected sib pairs (if available) helps reduce the number of candidate gene variants, even in consanguineous families. For X-linked recessive diseases, the analysis of the most remotely related male family members is the most effective strategy.
B. The trio-approach utilises data from unaffected parents and affected children, allowing for monitoring *de novo* variants.
C. The identification of *de novo* mosaic disease-causing variants requires a comparison of sequence data from affected and unaffected tissues in the same individuals.

2.7.2 Estimation of variant deleteriousness

Given the large output of WES and WGS, the clinical interpretation of genetic variants represents a crucial stage of the analytic process. Recent studies have indicated an unexpectedly large number of erroneously reported disease-causing mutations. Twenty-seven per cent of 460 causal variants cited were instead common polymorphisms or sequencing errors, or they lacked evidence of pathogenicity (Bell et *al.* 2002). More recently, a large-scale population study showed that on average, WES can retrieve ~54 variants in all individuals previously recorded as disease-causing variants. Most of these inconsistencies are due to classification errors in the literature and/or catalogues or databases of human variations (Lek et *al.* 2016).

In an effort to improve and standardise the quality of clinical interpretations of genetic variants, the American College of Medical Genetics and Genomics (ACMG) released specific guidelines (Richards et *al.* 2015). A series of criteria (inclusive of gene- and population-level information and *in silico* prediction tools) can be used to portray the pathogenicity of variants based on the structure

and the evolutionary conservation of the affected protein/domain/residue. The resulting five-level scale ranges from benign to pathogenic, with the variants of unknown significance being among the most difficult to interpret.

According to these criteria, the most straightforward interpretation follows the identification of reported disease-associated variant(s). To establish a causal relation, the patient's phenotype and the inferred inheritance patterns of the candidate variant must match those associated with the gene. The interpretation process is more complex if the retrieved variant(s) are not previously reported but are in known disease genes. In this case, the overlap of phenotypic features with what is reported in the literature or the segregation data is not sufficient to confirm the genetic findings as causal. *In silico* programmes can be used to support the potential causative role of the variants and to prioritise candidate pathogenic novel variants. Nonetheless, they can't be considered a replacement for functional validations and have not yet achieved enough accuracy to be used in a clinical setup (Young et *al.* 2015; Seifi et *al.* 2018). Therefore, establishing whether an identified genomic variant could be causative or could determine medically actionable conditions remain a complex task without functional analyses.

2.8 Primary Immunodeficiency

2.8.1 Overview

Primary Immunodeficiency (PIDs) comprise a large and heterogeneous spectrum of monogenic germline disorders that affect the development and/or functions of the immune system. Despite being generally quite rare, the global frequency of PIDs is estimated to be around 1:1200 live births, and the reported incidence is likely underestimated due to the difficulties in diagnosing patients with atypical clinical presentations (Bousfiha et *al.* 2013; Seleman et *al.* 2017). The classification of PIDs is generally based on their clinical characteristics. Usually, an initial clinical categorisation of PID suspects is proposed along with targeted screening tests to determine any likely causative genetic defects to select diagnostic therapeutic strategies (Seleman et *al.* 2017). Nonetheless, the presented approach can be time-consuming and expensive, especially if applied to heterogeneous and/or elusive phenotypes.

The introduction of NGS technologies has allowed for targeting multiple genes in synchrony, setting the stage for the development of extensive molecular diagnostic tests (Seleman et *al.* 2017). Due to the emergence of NGS applications in diagnostics and research, the number of recognised PIDs swiftly increased. To date, alterations in 344 different genes have been reported in association with more than 350 distinct conditions (Picard et *al.* 2015, 2018). Due to improved molecular diagnostics, the phenotypic spectrum associated with several individual diseases was extended. The broad variability identified in many individual PIDs reflect allelic heterogeneity and/or host/environmental modifying factors (Picard et *al.* 2015).

The spectrum and prevalence of PIDs in Finland have only been partially characterised. Recently, a cross-sectional study depicted an unprecedented high prevalence of Common Variable Immunodeficiency (CVID) in the Finnish population. CVID represents the most prevalent PID worldwide (2-4/100000), and the enrichment in Finland (6,9/100.000) is likely determined by the unique feature of the population (Selenius et *al.* 2017).

The population history of Finland has in fact led to a phenomenon defined as the Finnish Disease Heritage (FDH), distinguished by the enrichment of some detrimental disease-causing variants and a loss of others. The resulting FDH disorders are more frequent in Finland than elsewhere. For each of the FDH disorders, most (or all) of the affected individuals are identified by a causative founder mutation (Fin_{major}), displaying allelic frequencies higher in Finland than in other populations (Peltonen et *al.* 1999; Norio et *al.* 2003). Thus far, three PIDs have been included in the FDH: cartilage-hair hypoplasia (CHH), autoimmune polyendocrinopathy (APECED) and Cohen syndrome (Neufeld et *al.* 1981; Makitie et *al.* 1992; Ridanpaa et *al.* 2001; Norio 2003; Polvi et *al.* 2013). The currently available population-level exome data show two Fin_{major} mutations significantly more frequent in Finns than in other populations with European ancestries. The mutations cause APECED (rs121434254, 6.25-fold enriched) and Cohen syndrome (rs180177327, 47.11-fold), respectively. No reliable data are available for CHH-associated variants.

2.8.2 Disease classification

The classification of PIDs is performed, revised and implemented since 1990 by the Primary Immunodeficiency Expert Committee (PID EC) of the International Union of Immunological Societies (IUIS) and joint task forces,), now named Inborn Errors of Immunity Committee (Al-Herz et *al.* 2014; Bonilla et *al.* 2015; Picard et *al.* 2015, 2018, Bousfiha et *al.* 2018). These catalogues of inborn errors of immunity served as a reference for immunologists and researchers worldwide. The most updated phenotypic classification of PIDs is listed in Table 2.

I. Immunodeficiencies affecting cellular and humoral immunity
Caused by defects in T- and B-cells, these immunodeficiencies most often cause increased susceptibility to infections (bacterial or viral), at times with autoimmune manifestations.
II. Combined immunodeficiencies with associated or syndromic features
Conditions with distinguished phenotypes inclusive of immune defects.
III. Predominantly antibody deficiencies
This group accounts for approximately half of all patients with a PID diagnosis.
IV. Diseases of immune dysregulation
These diseases are caused by a defective control of the adaptive immunity, often characterised by autoimmunity and/or haematological malignancies.
V. Congenital defects of phagocyte number or function
These defects are caused by a severe reduction in the number or function of phagocytes.
VI. Defects in intrinsic and innate immunity
This category includes conditions determined by defects in components of the innate immunity, usually associated with isolated susceptibility to specific pathogens.
VII. Autoinflammatory disorders
This is a group of conditions causing (recurrent) febrile episodes characterised by inflammatory manifestations (cutaneous, mucosal, serosal and osteoarticular) (Fietta et *al.* 2004; Zeft et *al.* 2012; Ozkurede et *al.* 2012).
VIII. Complement deficiencies
These deficiencies are associated with varied clinical manifestations related to the altered branch(-es) of a pathway.
IX. Phenocopies of inborn errors of immunity
These errors are diagnostic relevant non-inborn errors of immunity resembling inherited PIDs (Picard et *al.* 2018).

Table 2. Revised 2017 phenotypic classification of inborn errors of immunity.
Based on the accompanying 2017 IUIS Inborn Errors of Immunity Committee classification (Bousfiha et *al.* (2018)

2.8.3 Autosomal recessive hyper-IgM syndrome type 2

Hyper-IgM syndromes (HIGM) are rare PIDs characterised by early-onset recurrent infections with autoimmunity, also referred to as Immunoglobulin class switch recombination deficiencies (CSR-D). The typical immunological trait of HIGM type 2 (HIGM2) is the absence or very low levels of IgG, IgA and IgE along with elevated or normal serum IgM levels (Gulino et *al.* 2003). The clinical hallmark of HIGM is lymphadenopathy, mostly in cervical and mesenteric lymph nodes as described in the majority of patients (75%). Histological sections of the massively enlarged lymph nodes portray a notable follicular hyperplasia with giant germinal centres (GC) (Zeft et *al.* 2012). The phenotypes typically reflect causal defects in some of the proteins involved in class switch recombination (CSR) causing a flawed process of somatic hypermutation (SHM)

in the Ig variable (V) region, which represents a crucial stage for antibody (Ab) functionality (Figure 9). CSR-Ds are genetically heterogeneous due to the possibility of alterations in different pathways related to antibody maturation (e.g. B cell activation, the formation of GC, induction of CSR, CSR and SHM processes) (Durandy et *al.* 2013).

Known established Ig-CSR deficiency causing genes include *CD40, CD40LG*, *AICDA* and *UNG* (Durandy et *al.* 2007). CD40L/CD40-deficiencies feature defects in cellular immunity, determining susceptibility to opportunistic and viral infections that are not dependent on immunoglobulin responses (Revy et *al.* 1998). For other CSR-Ds, recurrent bacterial infections, mostly in the gastrointestinal and respiratory tracts, are instead determined by humoral immunity defects caused by altered CSR processes. The most common intrinsic B cell defects in CSR-D are caused by mutations in *AICDA* gene encoding for activation-induced cytidine deaminase (AID), which is estimated to affect fewer than $2/10^7$ individuals (Durandy et *al.* 2013; Revy et *al.*1998; Minegishi et *al.* 2000).

The AID function is crucial for the correct implementation of CSR, SHM and central B cell tolerance (Kuraoka et *al.* 2011; Meyers et *al.* 2011; Cogne et *al* 2013). In addition, AID participates in the removal of epigenetic signatures through active de-methylation (Kumar et *al* 2013). AID-deficient patients often present with autoimmunity, increasing the burden of life-threatening manifestations in HIGM patients (Quartier et *al.* 2004; Durandy et *al.* 2013). Despite the genetic heterogeneity of HIGM, AID and uracil DNA glycosylase (UNG) deficiencies cause similar immunologic phenotypes. The most relevant feature is the absence of CD19+CD27+IgD-IgM-switched memory B (smB) cells in the blood, while marginal zone CD19+CD27+IgD+IgM+ B (MZB) cells are normal or high. The recognisable features associated with AID and UNG deficiencies make them clearly detectable using screening techniques.

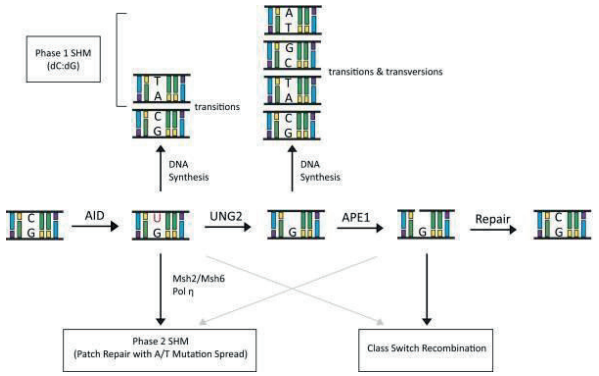

Figure 9. AID cytidine deamination activity in B cells. Adapted from Keim et *al.* (2013).
AID activity causes the deamination of cytidine residues. The resulting uracil residues are recognised for repair by the DNA base excision repair [BER] enzyme uracil DNA glycosylase (UNG) or by mismatch repair [MMR] factors (MSH2/MSH6). Based on the activity of error-prone DNA polymerases (DNA pol η), adjacent A/T-based pairs may be mutated to T/A, G/C or C/G base pairs. The uracil residues can also be excised by apurinic/apyrimidinic endonucleases (APE1 and APE2), causing single-strand (ss) DNA nicks. ssDNA nicks on both strands can generate DNA double-strand breaks (DSBs) during the class switch recombination (CSR) process.

2.8.4 Deficiency of adenosine deaminase 2

In 2014, biallelic deleterious *ADA2* (formerly *CECR1*) mutations were identified in patients with rare systemic autoinflammatory conditions characterised by vasculopathy and mild immunodeficiency defined as a deficiency of Adenosine deaminase 2 (DADA2). The vasculopathy symptoms in DADA2 patients featured primarily early-onset polyarteritis nodosa (PAN) and intracranial haemorrhagic and ischemic strokes. The chronic and recurrent systemic inflammation manifested with fever (with possible different skin phenotypes) and higher levels of acute phase reactants. The immunodeficiency symptoms are mostly similar to hypogammaglobulinemia. Overall, the occurrence of ischemic or haemorrhagic cerebral strokes represents the most severe manifestation and the primary cause of death for some DADA2 patients (Zhou et *al.* 2014; Elkan et *al.* 2014). In a zebrafish model, the transient knock-down of the orthologous of the human *ADA2*gene caused intracranial bleeding and severe neutropenia in embryos. The phenotype was prevented by a co-injection with wild-type human *ADA2* mRNA (van Montfrans et *al.* 2016). Subsequent studies added common variable immunodeficiency (CVID), Blackfan-Diamond anaemia, immune dysregulation, spastic paraplegia and bone marrow symptoms to the phenotypic spectrum of DADA2 (Table 3, van Montfrans et *al.* 2016; Hashem et *al.* 2017; Jain et *al.* 2017; Michniacki et *al.* 2018). Among the 43% of the patients with haematological manifestations, cytopenia symptoms were the most frequent (Table 3).

Currently, around 150 DADA2 patients have been reported in the literature (Table 3), some following clinical DADA2 suspicion while other after WES. All affected patients have been deficient in the plasma activity of the ADA2 protein. A large variability in clinical presentation and a lack of a clear correlation with *ADA2* genotypes emerged from the reported cases (Figure 10).

Interestingly, another ADA protein (ADA1) has previously been described in humans, which is encoded by the *ADA* gene. Both ADAs present adenosine deaminase activity (Franco et *al.* 2007; Zavialov et *al.* 2010). Mutations in *ADA* have been shown to cause severe combined immunodeficiency (SCID) (Buckley 2004). SCIDs are caused by the intracellular accumulation of toxic adenosine products resulting in apoptosis, depleting the count of circulating T, B and NK lymphocytes (Franco et *al.* 2007). Despite the similarities, human adenosine deaminases bind to distinct lymphocyte subsets that express specific receptors for ADA1 and ADA2 (Kaljas et *al.* 2017). In addition to differences in deaminase functions, ADA2 is secreted by activated monocytes and determines their proliferation and macrophage differentiation (Zavialov et *al.* 2010). Thus, a possible role of ADA2 in the balance between pro-inflammatory (M1) and anti-inflammatory (M2) monocytes has been postulated. In support of this postulation, experimental silencing of *ADA2*in myeloid cells resulted in a reduction of monocytes to macrophage differentiation and a higher prevalence of pro-inflammatory (M1) cells (Zhou et *al.* 2014).

Diverse therapeutic options have been suggested for DADA2, usually depending on the variety and severity of the symptoms in individual patients. Corticosteroids are used for the temporary relief of symptoms, and only anti-TNF agents among various immunosuppressants have been shown to lead to disease remission in patients with or without vascular lesions (Schepp et *al.* 2017).

Haematopoietic stem cell transplantation (HSCT) can be conceived as a potential definitive option in DADA2 (Van Eyck et *al.* 2015), even in patients with severe haematological and/or immunological phenotypes or who are unresponsive to vasculopathy treatment (Hashem et *al.* 2017).

Report	Patients with hematological manifestations/total patients	Hematologic manifestations	BM findings
Zhou et al. (2014)	6/9	Leukopenia (n=3); Pancytopenia (n=2); Evans syndrome (n=1)	BM myelofibrosis and lymphohistiocytic aggregates (n=1)
Elkan et al. (2014)	0/24	-	
Garg et al. (2014)	0/1	-	
Belot et al. (2014)	0/2	-	
van Montfrans et al. (2014)	2/2	Lymphopenia (n=1); Granulocytopenia and Anemia (n=1)	
van Eyck et al. (2014, 2015)	1/2	Cytopenia and lymphoproliferation	
van Eyck et al. (2014)	1/1	Anemia	
Bras et al. (2014)	0/5	-	
Westendorp et al. (2015)	0/2	-	
Gonzalez Santiago et al. (2015); Abraam et al- (2016)	0/2	-	
Batu et al. (2015)	2/6	Macrophage activation syndrome (n=1).	B-cytopenia additionally with myelofibrosis (n=1)
Schepp et al. (2016)	2/2	Anemia (n=1)	
van Montfrans et al. (2016)	9/9	Lymphopenia and anemia (n=9); thrombocytopenia (n=3)	
Fellmann et al. (2016)	1/2	Neutropenia	
Ben-Ami et al. (2016)	5/5	Anemia (n=5); thrombocytopenia and lymphopenia were additionally described	Red cell aplasia on BM evaluation (n=2)
Keer et al. (2016)	0/1	-	
Hsu et al. (2016)	1/1	Patient developed BM failure (neutropenia and hypocellular marrow) necessitating BM transplantation.	Lymphohistiocytic aggregates noted on marrow analysis.
Poswar et al. (2016)	0/1	-	
Nanthapisal et al. (2016)	6/15	Lymphopenia (n=6)	
Uettwiller et al. (2016)	2/2	Lymphopenia (n=1); lymphopenia and anemia (n=1)	
Elbracht et al. (2017)	0/1	-	
Hashem et al. (2017)	9/9	5 patients previously reported. pancytopenia (n=4); Pure red cell aplasia (n=6); neutropenia (n=2); immune thrombocytopenia (n=1); anemia and lymphopenia (n=1)	Pre-transplant marrow pathology for patients not described in report.
Hashem et al. (2017)	1/1	BM failure (pure red cell aplasia and neutropenia); treated with BM transplantation	Erythroid hypoplasia, decreased granulopoiesis, mild eosinophilia and increased T-lymphocytes with normal cellularity.
Omoyinmi et al. (2017)	0/2		
Caorsi et al. (2017)	0/15	-	
Skrabl-Baumgartner et al. (2017)	2/2	Anemia (n=2)	
Schepp et al. (2017)	7/11	2 patients previously reported. Lymphopenia and anemia (n=3); lymphopenia (n=1); lymphopenia and thrombocytopenia (n=1); anemia and thrombocytopenia (n=1); neutropenia and thrombocytopenia (n = 1)	
Alsultan et al. (2017)	1/1	Lymphadenopathy; thrombocytopenia	
Lamprecht et al. (2018)	0/1		
Cipe et al. (2018)	1/1	Neutropenia	
Gunthner et al. (2018)	0/1		
Sahin et al. (2018)	4/8	2 patients previously reported; lymphopenia (n=3); pancytopenia (n=1)	
All novel *DADA2* patients with hematological manifestations	63/147 (42.8%)	Patients described in more than one publication accounted for in prevalence calculation	

Table 3. Published reports of patients with DADA2 and the occurrence of haematological manifestatio

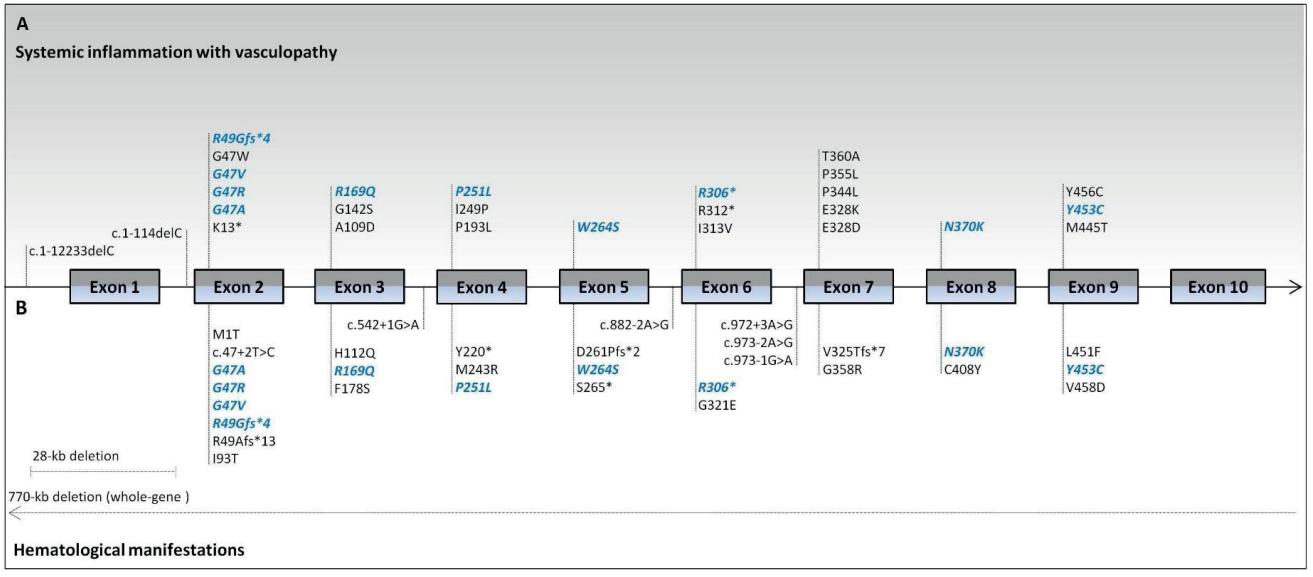

Figure 10
Deficiency of adenosine deaminase 2 (DADA2)-causing variants described in published reports until 2018.
Panel A shows the variants reported in patients with the autoinflammatory/vasculitic phenotype.
Panel B shows the variants reported in patients with hematological manifestations. The variants occurring in both sections (blue) are listed in italic.
The variants in the *ADA2* gene refer to the next reference sequence and genomic location: RefSeq NM_001282225; chr22:17659680-17700475 (GRCh37.p13).

2.8.5 Telomere biology disorders

Telomeropathies or telomere biology disorders (TBDs) consist of a range of clinical conditions resulting from impaired telomerase function. Telomerases are enzyme complexes specialised for the retention of terminal chromosomal DNA sequences, namely telomeres, which are not duplicated during the DNA replication process. Telomeres consist of repetitive DNA sequences located at each end of the chromosomes that protect the extremities from degradation or spontaneous DNA-repair activities. In humans, telomeres consist of tandem repeats of the 'TTAGGG' motive (highly variable in length among different individuals), which is the substrate for telomerase activity (Figure 10; Sandin et *al.* 2014; Bertuch 2016).

The first inherited disorder described as a TBD was dyskeratosis congenita (DKC). DKC is a severe multisystem disease presenting with cytopenia of one or more haematopoietic cell lineages with the clinical hallmarks of reticulate skin pigmentation, nail dystrophy and oral leucoplakia (Heiss et *al.* 1998). The most serious clinical feature in DKC is bone marrow failure, which is detected in the majority of patients and determines premature mortality. Furthermore, DKC patients display a heightened risk for malignancies, fatal pulmonary complications and immunodeficiency (Bertuch 2016). Currently, DKC is associated with alterations in 11 genes: *DKC1* (causing X-linked recessive inheritance), *TERT, TERC, NHP2, NOP10, ACD, WRAP53*, *TINF2, RTEL1, CTC1* and *PARN*. All these genes except *DKC1* are associated with autosomal recessive and/or dominant inheritance (Dokal 2000; Vulliamy et *al.* 2006; Ballew et *al.* 2013; Townsley et *al.* 2014; Bertuch 2016). According to recent classifications, the term 'DKC' is mostly used to classify well-defined childhood symptoms (Townsley et *al.* 2014). Based on clinical signs, a rare, severe phenotypic variant of DKC is defined as Hoyeraal-Hreidarsson syndrome (HHS). HHS leads to immunodeficiency, cerebral hypoplasia and intra-uterine growth retardation in addition to bone marrow failure (Dokal 2000; Glousker 2005).

From a biological perspective, the premature senescence of stem cells causes telomere shortening, which is markedly evident in the highly proliferating mucocutaneous tissues (Bertuch 2016). In the absence or delayed onset of pathognomonic mucocutaneous findings, the diagnosis of DKC has often been overlooked (Lim et *al.* 2014). Molecular diagnostic approaches based on NGS could improve the detection rate of TBDs, which is difficult due to the wide spectrum of clinical presentations and the lack of a conclusive laboratory test. The value of a prompt diagnosis is evident, especially considering the limited span of available therapeutic options for TBDs.

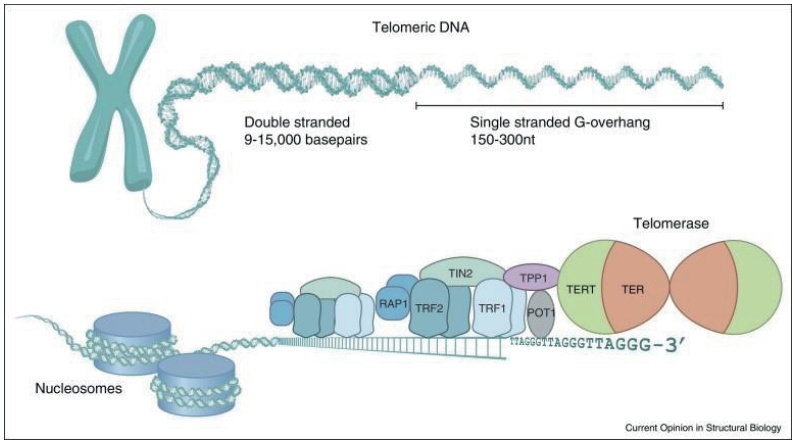

Figure 11. Structure of human telomeres and telomerase recruitment. Adapted from Sandin and Rhodes (2014).
Telomeres consist of repeats of the 'TTAGGG' motive, extending from the telomeric extremity to constitute a long double-stranded DNA region ending with a single-stranded G-rich overhang. In addition to histone-mediated packaging, telomeric DNA is organised into a shelterin capping structure by the binds with sequence-specific DNA binding proteins (TRF1, TRF2, Rap1, TIN2, TPP1, and POT1). TRF1 and TRF2 bind to specific double-stranded telomeric DNA sequences, and TIN2 allows for the formation of bridges with TRF1 and TRF2 and TPP1. The TPP1/POT1 complex is located on the G-overhang through the binding of POT1 with a single-stranded G-overhang. The telomerase complex TERT/TER begins its activity from the end of the telomeres. The recruitment requires an interaction between TPP1 and TERT (N-terminal domain) and specific sequence matching between the G-overhang and the template region of TER (telomerase RNA subunit).

2.9 Population structures of Finns and the Finnish disease heritage

Currently, the availability of a large-scale variation dataset allows for the characterisation of the genetic features of different populations and the link with clinical phenotypes. Demography can shape the genetic structure of populations and can result in different variation loads among individuals. A specific example is provided by the proportional increase of detrimental alleles in European American individuals compared to African Americans due to the so-called 'Out-of-Africa bottleneck' (Lohmueller et *al.* 2008). Conversely, other studies depicted similar loads of deleterious alleles in Europeans and West Africans, suggesting that demographic factors are unlikely to affect the proportions of deleterious variants in human populations or the average burden of mutations in individuals (Simons et *al.* 2014).

The actual Finnish population has been considered to likely be derived from small founder groups settling in the country after the glacial course. Consequently, the population history of Finland includes a restricted number of founders, isolation, several population bottlenecks and recent expansions (Peltonen et *al.* 1999). A comparison to non-Finnish Europeans showed a proportional enrichment of rare deleterious variants in Finns, notwithstanding the notably low rate of overall variations in the population. As assumed, the distribution of common alleles (non-synonymous or synonymous) appears to be unvaried. Instead, the frequency of rare detrimental variants has increased due to the founding bottleneck processes, and there was possibly not enough

time for the negative selection to reduce the load. The increased load of rare deleterious variants may also have resulted due to aspects of the Finnish population's shaping processes other than bottleneck processes (Lim et *al.* 2014).

The noticeable enrichment of some deleterious variants and the loss of others resulted in a phenomenon called the Finnish Disease Heritage (FDH). According to the definition, FDH disorders are more frequent in Finland than in
any other country (if present). For each FDH disorder, a causative founder mutation (Fin_{major}) accounts for the majority, if not the total, of affected individuals, with an allelic frequency higher in Finland than elsewhere (Peltonen et *al.* 1999; Norio 2003). At present, a total of 36 disorders is included in the FDH with reported mutations in 40 genes (Polvi et *al.* 2013; Kääriäinen et *al.* 2017) (Figure 11). The overall impact of FDH must not go underestimated due to the low prevalence of the individual disorders (from 1:8000 to 1:100.000 in Finland). FDH syndromes present a wide spectrum of associated phenotypes. Estimates report one-third of the conditions associated with (mild to severe) intellectual disabilities, one third with vision defects and one-half causing premature death (Norio 2003). In addition to the peculiar population architecture, genetic research in Finland revealed the possibility to access detailed genealogical records and longitudinal national health registers, allowing for the development of personalised risk evaluation strategies in preventive healthcare (Kääriäinen et *al.* 2017).

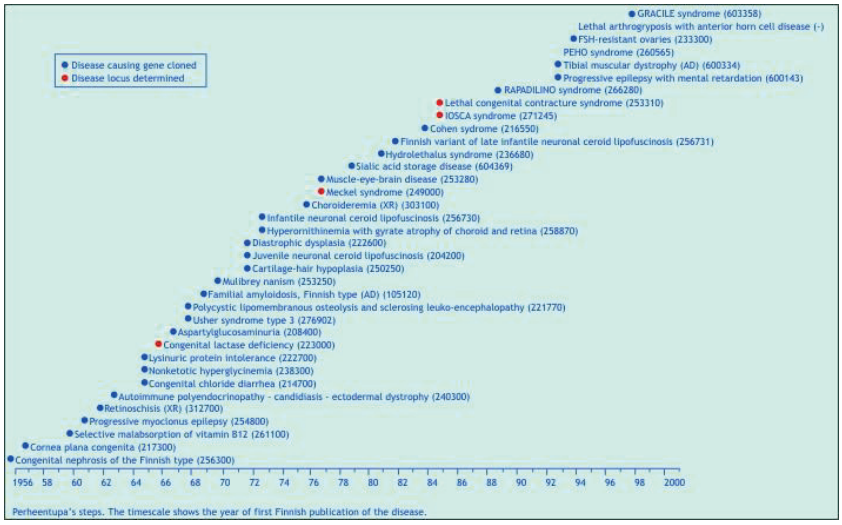

Figure 12. List of diseases of FDH (as 'Perheentupa's Steps'), from http:// http://findis.org/.

3 AIMS OF THE STUDY

The purpose of this thesis study was to shed light on the genetic bases of PIDs in the Finnish population and to implement NGS-based methods for routine diagnostic testing.

Part of the study subjects had previously been ruled out for alterations of established candidate disease-genes. Moreover, the varied clinical manifestations within the cohort and the overlap of symptoms among different syndromes would have prompted the possibility of pathogenic variations in a large number of causal-genes. Considering these hurdles, we targeted all the known protein-coding genes, without setting clinical-based assumptions or restrictions. This, namely hypothesis-free, WES-based approach was applied to identify PID-causing variants and to provide prompt genotype-phenotype correlations for heterogeneous syndromes.

In detail, the goals of the project were:

1. To identify disease-causing genetic variants in Finnish patients with unexplained immune and haematological diseases using hypothesis-free WES-based methods.

2. To characterise the phenotypic spectrum of rare PID entities providing insights on the biological function of the involved genes.

3. To improve routine PIDs genetic diagnostics at the single patient-level, especially for highly heterogeneous and/or atypical clinical presentations.

4. To evaluate the effects of the genetic background of the Finnish population on the distribution of rare PID-causing mutations.

4 MATERIALS AND METHODS

4.1 Study subjects

All the genetic analyses performed in this study were intrinsic to clinical setting for a set of 212 Finnish individuals with a clinical suspicion of PIDs. The study subjects did not present consanguineous origins and displayed a broad spectrum of immunological and/or haematological phenotypes suggestive of monogenic disorders. Inclusion in the study was based on clinical diagnosis and/or laboratory findings performed at the University Hospitals of Finland (Helsinki, Kuopio, Oulu, and Tampere).

4.1.1 Ethical issues (I, II, III)

The studies were undertaken in observance of the principles of the Helsinki Declaration with the approval of the ethics committees of Helsinki and Oulu University Hospitals, Finland. All participants provided written informed consent preceding the sample collection. For the children enrolled in the study, written informed consent was provided by one of the parents.

4.2 Generation and analysis of genetic data

4.2.1 Whole-exome sequencing (I, II, III)

The WES was performed for all studied patients at the FIMM sequencing centre (the Institute for Molecular Medicine Finland, FIMM; Helsinki, Finland) with the exception of P1.1 (study III), who was tested at the Laboratory of Human Genetics of Infectious Diseases, Necker Branch, (Necker Hospital for Sick Children, Paris, France). In brief, genomic DNA and RNA were isolated from peripheral blood by standard methods. The SureSelect Clinical Research Capture Exome (Agilent, Santa Clara, CA, USA) and Nextera®Rapid Capture Exome (Illumina, San Diego, CA, USA) kits were used for target enrichment. Paired-end sequencing was carried out on the HiSeq 1500 or HiSeq 2000 platform (Illumina, San Diego, CA, USA). An in-house pipeline (VCP) was used for the analysis of the generated WES data (Sulonen *et al.* 2011).

4.2.2 Sequence read processing (I, II, III)

Initially, raw Illumina reads were merged (SeqPrep 0.4.5). The ends of the paired reads underwent trimming with all the resulting reads removed if shorter than 36 bases. The remaining paired and single reads were aligned separately against the human reference genome hg19 (GRCh37, *ensembl* version 70) using the Burrows Wheeler Aligner (0.5.10). The alignment of indels was enhanced by GATK Indel Realignment (1,5-3) (McKenna *et al.* 2010, DePristo *et al.* 2011, Van der Auwera *et al.* 2013). The results of the alignment were refined by the removal of potential PCR-related

biases (i.e. duplicates) (Picard MarkDuplicates, 1.65) and read pair mapping to multiple genomic targets (with one or both reads).

4.2.3 Single-nucleotide and indel variant calling (I, II, III)

The variant calling processes were carried out using SAMtools mpileup (0.1.18) and dindel (1.01). The minimum depth set for SNP calling was seven reads. Any call with a coverage ratio between reference (R) and variant (V) calls: (R / (R + V)) smaller than 0,2 was assumed to be homozygous, and the remainder was considered heterozygous. Quality filtering was applied to the called SNPs, discarding those with a ratio between the sum of the quality values of reference (R) and the variant (V) calls: R / (R + V) higher than 0,8. For two different variant calls and no reference for a given position, the highest in terms of the quality call was used as a reference.

4.2.4 Variant and gene annotation (I, II, III)

The Annovar toolkit was used to annotate and to predict the functional consequences of the identified variants (Wang *et al.* 2010, Chang *et al.* 2012).

The allelic frequencies in the general population were retrieved using the following databases: the national *SISu* project (http://sisu.fimm.fi/), the Exome Aggregation Consortium (ExAC, accessed in July 2015) (Lim *et al.* 2014) and the Genome Aggregation Database (gnomAD) (Cambridge, MA, USA; http://exac.broadinstitute.org; accessed in July 2015) (Lek *et al.* 2016).

The annotations of genes or variants to human phenotypes were based on the Online Mendelian Inheritance in Man, OMIM® database (https://omim.org/) (Hamosh *et al.* 2005), the NCBI ClinVar database (https://www.ncbi.nlm.nih.gov/clinvar/), the Human Gene Mutation Database (HGMD®) (Krawczak and Cooper, 1997), the LOVD database (http://databases.lovd.nl/shared/genes/) and the literature investigation.

4.2.5 Variant filtering and analysis (I, II, III)

To reduce the list of variants for the analysis, all common (frequencies above 0,01 in the general population) and non-coding variants were discarded. Filtering was based on allelic frequency data from gnomAD and *SISu*. Furthermore, all variants that exceeded the frequency of 0,05 (data from an in-house database) were rejected.

All rare variants located in the coding regions were prioritised based on the predicted transcript level consequences. Frameshift, nonsense, splicing and missense variants were included. For missense variants, the predicted effect on the protein level was estimated *in silico* by CADD (Kircher *et al.* 2014), PolyPhen-2 (Adzhubei *et al.* 2010), SIFT (Kumar *et al.* 2009) and Mutation Taster (Schwarz *et al.* 2014).

For multi-case families or a suspicion of family history for a specific disease, the phenotypic segregation pattern was taken into account during the filtering process.

4.2.6 Variant pathogenicity estimation (I, II, III)

All filtered variants were prioritised based on *in silico* prediction tools (prediction of the effect on the protein and the conservation of the affected amino acids) and underwent evaluation according to the ACMG Standards and Guidelines (Richards *et al.* 2015). Only variants classified as pathogenic, likely pathogenic or associated with an unknown significance were considered to have potential clinical relevance.

4.2.7 Variant confirmation and segregation analysis (I, II, III)

The confirmation of all the identified variants and the screening of family members (when available) were performed by PCR amplification on genomic DNA followed by Sanger sequencing. The sequencing was performed using the BigDye Terminator Cycle Sequencing kit and capillary electrophoresis executed on an ABI-3730XL DNA Analyser (Applied Biosystem, Foster City, CA, USA). The resulting electropherograms were analysed using Sequencer 5.1 (Gene Codes Corporation, Ann Arbor, MI USA). Some of the identified causal variants were tested at the transcript level by Sanger sequencing of cDNA obtained from RNA using the SuperScript® VILO cDNA Synthesis kit (Life Technologies, Carlsbad, CA). The kit was used according to the manufacturer's instructions (0,5-2µg of RNA, 20µl of total reaction volume).

4.2.8 Exclusion of germline mutations in the *GATA2* gene (II)

The hypothesis of *GATA2* haploinsufficiency in patients P1 and P2 (study II) was tested by Sanger sequencing of cDNA (as described above). In brief, heterozygous SNVs (previously detected by WES) in the coding region of *GATA2* (Ensembl ENSG00000168610, ENST00000588969; GRCh37.p13) were targeted to compare zygosity at the transcript level. The presence of only one allele in the cDNA was considered to potentially reveal haploinsufficiency. Because P1 did not carry any heterozygous coding variant in *GATA2*, all reported haploinsufficiency-causing variants (coding or not) were investigated according to the Human Gene Mutation Database.

4.2.9 Exclusion of somatic mutations in the *STAT3* gene (II)

Somatic *STAT3* (Ensembl ENSG00000179348, ENST00000341105; GRCh37.p13) variants were screened by deep-amplicon sequencing in patients P1-4 and P6-7 (II) and in two healthy controls. The targeted sequencing of the exons 19, 20 and 21 (corresponding to the functional Src Homologue 2 [SH2] domain) was performed using the Illumina Miseq platform (Illumina).

4.3 Functional studies of the identified variants

4.3.1 DADA2 measurement (II)

Plasma ADA2 activity levels were measured in seven individuals. For patients P1, P2, P6 and P7, the measurement was performed on EDTA-treated plasma (expressed as mU/mL; reference values: 12,8mU/mL, SD 2,6 (min-max) in adults (>18years old), 27mU/mL, SD 2,3 (min-max) in children (<18y)) (Zhou *et al.* 2014). In the remaining patients (F2.3, P4.1 and P4.2), DADA2 activity levels

were assessed on filter paper (activity measured in mU/g of protein, reference values: 132,3mU/g, SD 51.1). The previously mentioned method was partially modified to be executed on filter paper (Ben-Ami *et al.* 2016). Two of the patients were not tested (P3 underwent an allogenic stem cell transplant in September 2016; P5 was deceased in 1990).

4.3.2 Telomere length analysis (III)

The relative telomere length (RTL) was determined in the studied patients using a quantitative PCR-based method described by Cawthon (2002). Minor modifications were applied to the described protocol. Briefly, the samples consisting of 17,5ng of DNA (from peripheral blood leukocytes) were analysed for separate telomere (TEL) and single copy gene (HBG) reactions using the ABI 7900HT platform (Applied Biosystems). Each sample was analysed in triplicate wells in 96-well PCR plates on two separate occasions. Cell-line DNA (CCRF-CEM) and control samples were included in all runs to monitor PCR efficiency and to derive a standard curve. The TEL/HBG (T/S) values were derived using the $2^{-\Delta Ct}$ method ($\Delta Ct = Ct_{TEL}-Ct_{HBG}$). The RTL values were obtained by dividing the samples' T/S values by the T/S value of a reference DNA. The RTL values of the patients were compared to normal controls (n=113; age=0-83 years).

4.3.3 T-cell sorting and clonality analysis (II)

Peripheral blood mononuclear cells from two patients with LGL lymphoproliferation (P1, P2) were sorted into CD8+ and either CD8- or CD4+ populations by flow cytometry (BD Biosciences Multitest, CD45, CD3, CD8 and CD4 antibodies). A fraction of the isolated cells was used to check the purity of the sortings (FACS Aria, Beckman-Coulter Immunotech). Potential clonal expansions were tested by a quantitative determination of the TCR Vβ repertoire of human T lymphocytes, which was performed using flow cytometry. The assay was executed on peripheral blood samples using the IO Test® Beta Mark TCR Vβ Repertoire Kit (PN IM3497, Beckman Coulter Immunotech). In brief, the kit consists of eight vials with mixtures of 24 different conjugated TCR Vβ antibodies encompassing approximately 70% of the normal human TCR Vβ repertoire. The manufacturer established normal means (with minimum and maximum values) derived from 85 normal blood specimens (the tubes detecting Vβ4, Vβ7,2 and Vβ13,2 have been tested on a different cohort of 46 normal specimens). Subsequently, the samples were stained with CD3, CD4 and CD8 antibodies (BD Biosciences) to further characterise the possible clonal expansions.

4.4 Analysis of the distribution of PID-causing variants

4.4.1 Enrichment analysis (I, II)

Publicly available data (ExAC and gnomAD) were used for the comparison of the frequency of PID-causing variants in different populations. The allelic distributions in Finns and individuals with European and non-European ancestries were compared using Pearson's chi-squared test (10^8 simulations, statistical significance $p<0,05$). All the data were retrieved and analysed using in-house designed IT procedures.

4.4.2 Population study (I)

The estimation of haplotype blocks potentially including the *AICDA* p.(Met139Thr) (c.416T>C; rs200858797) variant was performed on data from the SISu cohort (3316 unaffected, 11 p.(Met139Thr) carriers) using PLINK (1,07, http://pngu.mgh.harvard.edu/purcell/plink/; Purcell et *al.* 2007) and Haploview (Barrett et *al.* 2005).

Sixty-one common markers selected in a 2Mb window around the target variant (Minor allele Frequency-MAF > 5%, genotyping rate threshold >95%, not failing Hardy Weinberg equilibrium test) were used. The following options were applied: '--maf 0,05 --geno 0,05 --hwe 1e-5 --block --hap-freq' (PLINK) and 'solid spine of LD option' (Haploview).

The results were replicated using the same analytical protocol and options on a different cohort (6735 healthy individuals, 20 p.(Met139Thr) carriers) of three Finnish clinical and epidemiological sample collections (the Finnish Twin Cohort study, the National FINRISK Study and the Migraine Family Study). The samples were genotyped with Infinium HumanCoreExome-24 BeadChips (Illumina). For the segregation analysis, 113 common markers (MAF > 5%) in a 2Mb window surrounding the p.(Met139Thr) were selected.

According to the detected haplotype structure, a potential ancestral Finnish haplotype including the studied *AICDA* variant was monitored. Therefore, the data from the two different datasets defining the 1,1Mb region surrounding the p.(Met139Thr) were defined. The genotypes of the enclosed 80 markers were extracted in all the carriers (n=31) from the different datasets. In detail, the genotypes of 34 markers were available in the SISu cohort, 51 were available in the other collections and only five markers were genotyped in both datasets.

A cryptic relatedness analysis was also performed on the samples from the Finnish Twin Cohort study, the National FINRISK Study and the Migraine Family Study cohorts. The overall IBD sharing was compared among the p.(Met139Thr) carriers (n=20) and the remainder of the individuals using 132.250 genome-wide markers (MAF >5%, genotyping rate threshold > 95%, not failing Hardy Weinberg equilibrium test) in apparently unrelated individuals (piHat<0.05, r^2>0,5). To estimate the individual genome-wide identity by descent (IBD), the following PLINK options were applied: '--genome --maf 0,05 --geno 0,05 --hwe 1e-5 --min 0,01'. The average values of the 'observed IBD sharing' (piHat) were compared in carriers and non-carriers using a Welch's two-sample t-test (108 simulations).

4.4.3 Geographic distribution analysis (I)

The geographic distribution of the *AICDA* p.(Met139Thr) alleles in Finland was studied by tracking the familial HIGM2 samples and by using data from the SISu consortium, the Finnish Twin Cohort study, the National FINRISK Study and the Migraine Family Study. Overall, information was retrieved about the place of birth for 27 of the 32 known p.(Met139Thr) carriers. Fifteen of the carriers originated from Eastern Finland (Kuopio, Viipuri), five from Northern-Ostrobothnia, four from Northern Finland (Ranua) and three from the Helsinki area (currently representing a mixed population).

5 RESULTS AND DISCUSSION

5.1 Overview of the study patients

The subjects included in the presented studies underwent clinical WES due to a suspicion of immune and/or haematological inherited diseases. All 212 study subjects were Finnish and mostly unrelated. A summary of the clinical characterisation of the patients is listed in Figure 13.

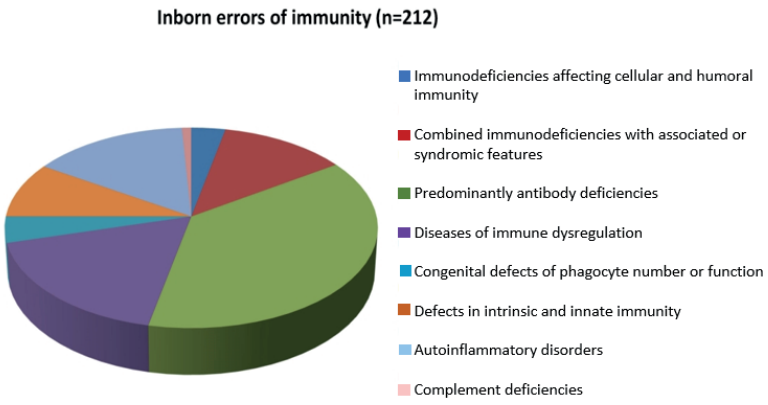

Figure 13. Categorisation of the clinical phenotypes of the patients in the studied dataset (n=212).
The categories were defined according to the 2017 Primary Immunodeficiency Disease Committee Report (Picard et *al.* 2018).

5.2 Identification of a disease-causing founder mutation in a population isolate (I)

5.2.1 Screening of *AICDA* mutations in Finnish patients with hyper-IgM syndrome type 2 (HIGM2)

The original index case displayed undetectable levels of switched memory B cells and other clinical features suggestive of familial HIGM2 because the other two family members presented similarly. All the affected members of the family (n=3) were remitted to WES, with an identification in all the patients of a bi-allelic HIGM2-causing mutation (p.(Met139Thr)), PMID:16964591) in the *AIDCA* gene (Ensembl ENSG00000111732, ENST00000229335, GRCh37.p13). In the case of a deceased familial patient (PI-I), targeted Sanger sequencing conducted on an archived sample retrieved the same genotype. The screening revealed two healthy relatives carrying one copy of the variant.

Following the identification of the disease-causing variant by WES, the patient cohorts of the paediatric and adult immunodeficiency units of all five Finnish university hospitals was screened

for phenotypes compatible with either AID or UNG deficiency. Patients with a) low or absent IgA, IgG and IgE levels but normal or high IgM levels according to the laboratory reference values along with b) missing smB cells but normal or high levels of MZB in B cell phenotyping (for methods, see Haapaniemi et *al.* 2015) were included in the screening. All the additional Finnish patients (n=4) were screened, and homozygotes were identified for the p.(Met139Thr) mutation.

5.2.2 Enrichment of AID-causing variant in the Finnish population

The HIGM2-causing p.(Met139Thr) mutation was found with an allelic frequency of 0,012% in the 57.391 exomes of the ExAC. The mutation is rarer in other population subsets included in the ExAC (allelic frequency of 0,0047% in non-Finnish individuals with European ancestry (n=31.686) and an absence of carriers in other populations of non-European ancestries (n=22.692)).

Comparing the different subgroups of European descent, the p.(Met139Thr) mutation emerged as largely enriched in Finland (38,56-fold allelic frequency, p <0,001) with 11 uniallelic carriers of 3013 exomes (MAF 0,18%). The same level of enrichment was confirmed using the gnomAD dataset. The analysis of the allelic frequency of other AICDA variants did not show statistically significant differences between different populations. Interestingly, no other deleterious *AICDA* mutations have been previously identified in Finnish AID deficiency patients, suggesting the p.(Met139Thr) is its causal factor in Finland. Thus, according to the p.(Met139Thr) allelic frequency, the theoretical prevalence of HIGM2 patients among Finns should be in the range of $0{,}81/10^6$ individuals.

5.2.2.1 Analysis of the geographic distribution of the founder allele

To study the geographical distribution of the enriched *AICDA* p.(Met139Thr) in Finland, information on the birthplace of the studied individuals and of all the carriers (27 out of 31) included in the SiSu dataset and in other Finnish sample collections was obtained.

All the HIGM2 patients and almost all the carriers (24/27) originated from Eastern and northeastern Finland. These regions namely correspond to the late settlement areas of the Finnish population, implying a single origin of the p.(Met139Thr) alleles that all of the patients carried (Figure 14). Three of the carriers were born in the Helsinki area, which during the last centuries has been characterised by a considerable immigration from the rest of Finland.

5.2.2.2 Identification of the founder mutation in the AICDA gene

To decipher the possible link between the enrichment of the p.(Met139Thr) and the genetic background of Finns, haplotypes shared by the carriers were identified using data from the SiSu cohort (3325 WES of Finns inclusive of 11 carriers.) Initially, the potential haplotype structure of a 2Mb genomic region around the p.(Met139Thr) was determined. Two haplotype blocks were noted 90 kb upstream and 51kb downstream of the variant, respectively. From a genomic perspective, a ‘recombination hotspot’ was observed spanning 10kb and enclosing the *AICDA* gene

(UCSC Genome Browser, Kent et *al.* 2002). It was speculated that the presence of a 'recombination hotspot' could have potentially reduced the conservation of the ancestral allele as well as the possibility to track it. Therefore, to increase the detection power, the genetic data retrieved for all the p.(Met139Thr) carriers present in the two different datasets were combined. The WES data from the SiSu cohort and the two familial carriers (exome sequenced as part of the clinical workout) and the genotyping data from 29 individuals of the Finnish epidemiological and clinical cohorts were also combined. Overall, for the 31 carriers, the different possible allelic combinations for each of the retrieved haplotype blocks were compared. A 207,5kb core haplotype inclusive of the p.(Met139Thr) variant present in all the carriers (Figure 15) was identified. The subsequent comparison of the genomic identity (measured as pairwise genome-wide IBD) showed a significant higher relatedness within the p.(Met139Thr) carriers (average piHat=0,007±0,0027) than in the general population (piHat=0,003±0,005)) (p=1,59e^{-12}). Although the structure of a potential founder haplotype was disrupted due to a recombination hotspot, the genomic identity and the haplotypes shared among the carriers were consistent with the single origin of the p.(Met139Thr) in Finns.

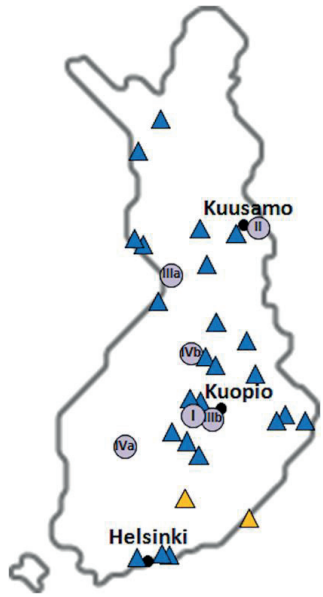

Figure 14. Distribution of the *AICDA* p.(Met139Thr) carriers in Finland.
Blue triangles point to the geographical origins of the Finnish carriers (n=27) of the p.(Met139Thr) variant included in the *SISu* and in the epidemiological and clinical Finnish sample collections (the Finnish Twin Cohort study, the National Finrisk Study and the Migraine Family Study). Yellow symbols indicate the birthplaces of carriers' parents if discordant. The birthplaces of the patients are indicated by a purple spot, listing the number of the Family (from I to IV). For Families III and IV, the mother corresponds to 'a' and the father to 'b'. The black dots mark the main municipal areas.

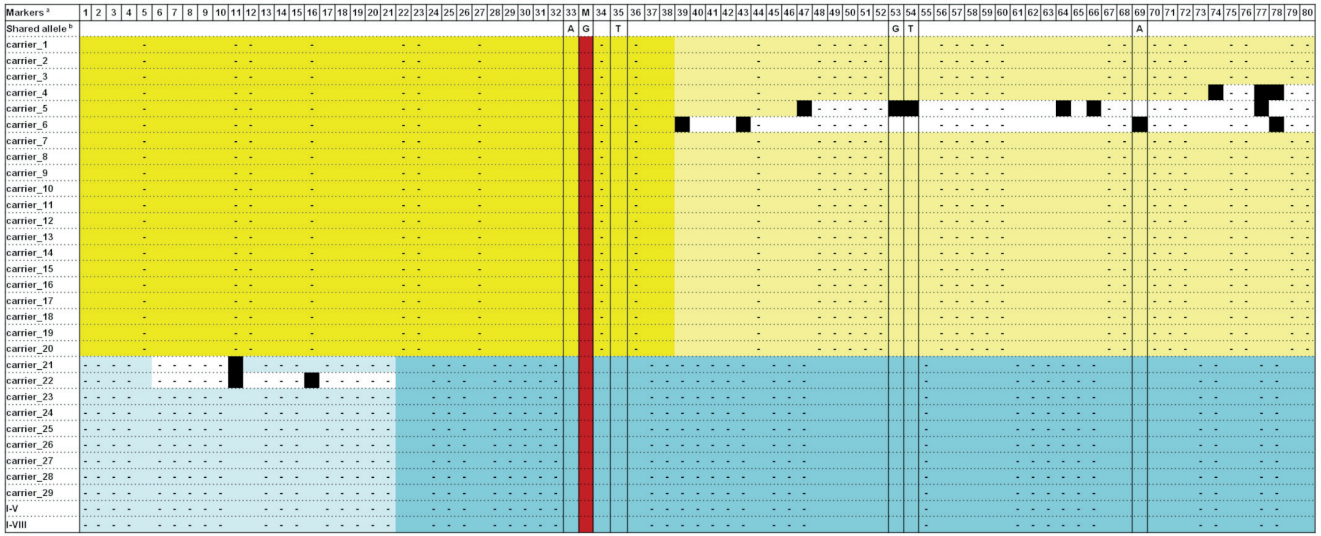

Figure 15. Haplotype structure of the flanking region of the *AICDA* gene in the 31 Finnish carriers of the p.(Met139Thr) variant.
The haplotypes of the carriers analysed by the genotyping chip (the Finnish Twin Cohort study, the National Finrisk Study and the Migraine Family Study) are shown on horizontal lines on a yellow background in the top part of the panel.
The haplotypes of the carriers analysed by WES (*SISu* project and study subjects of Family I) are presented on a blue background in the lower part of the panel.
The red column shows the position of the p.(Met139Thr) variant. Missing genotypes are marked by '- '. The yellow/blue squares show the identified shared haplotype in each mutation carrier, white filling indicates non-informative genotypes and black squares label recombination events (i.e. absence of the allele included in the above-mentioned haplotype). The minimum regions shared by all mutation carriers in each data set are indicated by the darker colour.

[a] The markers used in the analysis are indicated with numbers in the top row (marker names listed in Supplementary Table 3).

[b] The columns framed by black lines highlight the markers shared by both datasets, and the alleles that are seen in the shared haplotype are shown above.

5.3 Identification of novel phenotypic traits in patients with adenosine deaminase 2 deficiency (II)

5.3.1 Identification of *ADA2* mutations in Finnish patients with adenosine deaminase 2 deficiency (DADA2)

The sequencing of patients with a clinical suspicion of PIDs identified seven individuals with biallelic variants in the *ADA2* gene (Ensembl ENSG00000093072, ENST00000399837, GRCh37.p13). Part of the studied subjects presented with cytopenia (P1-P4.1), and others belonged to a multi-case family with cutaneous Polyarteritis nodosa (cPAN) and central nervous system complications (P5-P7) (Table 4). All the patients had non-consanguineous origins and carried the p.Arg169Gln mutation (PMID 24552285): in three cases, homozygotes (P1-P3), and in four cases, a compound heterozygote with pathogenic variants (the novel p.(Leu311Arg) in P4.1 and the c.973-2A>G (P5-P7). The p.(Leu311Arg) variant was defined as pathogenic according to the ACMG criteria, the c.973-2A>G according to the ClinVar database (ID541735). The analysis was extended to the patients' relatives by targeted screening, identifying two other patients (F2-3, sister of P2 and P4.2; aunt of P4.1).

5.3.1.1 Functional studies confirming DADA2

In seven patients, deficient levels of ADA2 protein activity (in the plasma) were functionally assessed. Patient P3 was not included in the testing because an allogeneic stem cell transplant was performed (September 2016). P5 had previously deceased (1990).

5.3.1.2 Enrichment of rare DADA2-causing variants in the Finnish population

The p.Arg169Gln has a frequency of 0,05% in the gnomAD dataset (of 138.611 genome-sequence d samples). The allelic frequency in Finns was 0,2% (52 carriers of 12.896 individuals), which is h igher than in the general population. Compared with the group of non-Finnish European individua ls (0,06% in 63.354 individuals), the p.Arg169Gln resulted in 3,32-fold enrichment in Finns (p <1 e^{-8}). The p.(Leu311Arg) variant has a frequency of 0,005685% in the gnomAD dataset and 0,049 % in Finns. The mutation resulted in a significant enrichment in Finns in comparison with other g roups of European descent (18,36-fold, p=5,5e^{-7}). The c.973-2A>G variant displayed a MAF of 0, 012% without significant differences among the population groups.

5.3.2 Characterisation of the Finnish patients with atypical phenotypes

The initially studied four patients (P1-P4.1) displayed variable degrees of cytopenia, and for three of them (P1-P3), signs of lymphoproliferation were described (Table 4). Interestingly, they did not

present any of the autoimmune vascular or haemorrhagic complications typical of DADA2. A clinical evaluation of the patients' relatives indicated potential lymphoproliferation symptoms in two patients (F2-3, P4.2). F2-3 was an asymptomatic sister of P2 presenting with mild cytopenia. P4.2 was an aunt of P4.1, originally diagnosed with CVID.

The absence of the canonical DADA2 phenotypes along with the recurrence of haematological manifestations creates additional options for the clinical management of the patients. The varied haematological (mostly cytopenia) and lymphoproliferative symptoms manifesting in several DADA2 patients might require targeted therapeutic options. After DADA2 is ascertained, the clinical strategies should be developed in consideration of the possibility of autoimmune or haemorrhagic manifestations even in absence of PAN or other vasculitic phenotypes.

5.3.2.1 Lymphoproliferative phenotype resembling T cell large granular lymphocyte leukaemia

Of the three patients with lymphoproliferation (P1-P3), two (P1, P2) exhibited a CD3+CD8+ T cell large granular lymphocytic (T-LGL) infiltration of the bone marrow. Autoimmune findings suggestive of T-LGL were first reported in P1 at the age of 8y and confirmed at 17y. For P2, the diagnosis of T-LGL was performed at 31y. P3 already showed lymphadenopathy and T-cell hyperplasia in the bone marrow at the age of 4y. P4.1 was only 5y at the time of the last clinical evaluation; therefore, given the variable onset of lymphoproliferation, the possibility of developing it later in life cannot be ruled out. F2.3 was initially reported to have mild cytopenia, but a further detailed immunological evaluation (age of 46) showed a marginal T-LGL clone in the peripheral blood (0,06x10-6/l), although it did not fulfil the standard criteria for a diagnosis of T-LGL leukaemia. P4.2 was initially diagnosed with CVID, but supplementary examinations at the age of 18 revealed mild signs of lymphoproliferation. Altogether, these findings indicate the occurrence of early-onset T-cell clonal expansion in patients with DADA2, though the causal link is still unclear.

5.3.2.2 T-cell clonality analysis

A T-cell clonality analysis was performed in the two patients with LGL lymphoproliferation (P1, P2). Through flow cytometry, peripheral blood mononuclear cells were sorted into CD8+ and either CD8- or CD4+ populations. The screening showed clonal expansions of the CD8+ fraction in P1 (Vb.5,1:18,8% and Vb.1:12,4%). Data are not available for P2.

ID	Sex	Age	*ADA2* genotype	Other genetic variation	Plasma ADA2 activity	Phenotype	Lymphoproliferation[f]
P1	F	43	p.Arg169Gln/p.Arg169Gln	Not detected	<0.25[b]**	Pneumonia, warts, hepatosplenomegaly, AIHA, neutropenia, thrombocytopenia, extremely low or non-existing dendritic, NK and B cells, T-LGL lymphoproliferation, chronic histoplasmose, hypogamma, PAH	Yes
P2	F	40	p.Arg169Gln/p.Arg169Gln	Not detected	<0.25[b]**	Anemia, hepatosplenomegaly, AIHA, T-LGL lymphoproliferation, extremely low or non-existing dendritic and NK cells, bone marrow lymphocyte infiltrate	Yes
F2-3*	F	46	p.Arg169Gln/p.Arg169Gln	Not detected	1[c]**	Low IgM, low NK-cells, low response to difteria vaccination.	very low
P3	F	5	p.Arg169Gln/p.Arg169Gln	15q26.1-q26.2 deletion	NA[d]	Severe neutropenia, lymphopenia, leukopenia, trombocytopenia, splenomegaly, lymphadenopathy, grade I myelofibrosis in bone marrow. Diamond-Blackfan anemia diagnosed as a newborn, recurrent fevers, developmental delay, dysmorphic features, epilepsy	Yes
P4.1	F	6	p.Arg169Gln/p.(Leu311Arg)	Not detected	0.3[c]**	Severe neutropenia, leukopenia, thrombocytopenia, splenomegaly, warts, atopy, and cellulitis	Not detected
P4.2	F	18	p.Arg169Gln/p.Arg169Gln	Not Detected	0.5[c]**	CVID, Raynaud syndrome, mild livedo reticularis and acrocyanosis, trombocytopenia, high B-CD19+ cells, very low vaccine responses, mild splenomegaly and para-aortal lymphadenopathy	Yes
P5	M	16.5[a]	p.Arg169Gln/c.973-2A>G	Not detected	NA[e]	cPAN, lymphopenia	Not detected
P6	F	38	p.Arg169Gln/c.973-2A>G	Not detected	<0.25[b]**	cPAN, lymphopenia	Not detected
P7	F	40	p.Arg169Gln/c.973-2A>G	Not detected	<0.25[b]**	cPAN, lymphopenia	Not detected

Table 4. Characteristics of DADA2 patients.
* asymptomatic sister of P2; ** deficient plasma ADA2 activity.
Abbreviations: F, female; M, male; NA, not available; Yes, present; AIHA: idiopathic autoimmune hemolytic anemia; EVANS, Evan´s syndrome; PAH, pulmonary arterial hypertension; CVID, common variable immunodeficiency; cPAN, cutaneous polyarteritis nodosa.
The variants in the *ADA2* gene refer to the next reference sequences: Ensembl ENSG00000093072, ENST00000399837 (GRCh37.p13). The genomic and cDNA location of the listed variants: p.Arg169Gln. (chr22:17687997C/T; c.506G>A.), p.(Leu311Arg) (chr22:17670872A/C; c.932T>G), c.973-2A>G (chr22:17669339T/C; c.973-2A>G).
[a]Age at death; [b]Reference values for plasma ADA2 activity: 12.8mU/mL, SD 2.6 (min-max) in adults (>18years old), 27mU/mL, SD 2.3 (min-max) in children (<18y); [c]Reference values for plasma ADA2 activity: 132.3mU/g, SD 51.1 (min-max); [d]Allogenic stem cell transplant was performed in 2016; [e]Deceased in 1990; [f]For more detailed clinical history see the Supplementary information.

5.3.2.3 Exclusion of germline *GATA2* variants in patients with T-LGL

The clinical phenotype of P1 and P2 partly overlapped with features of *GATA2* deficiency or with the phenotypes previously observed in patients carrying activating *STAT3* mutations (reported autoimmunity, hypogammaglobulinemia and lymphoproliferation). Therefore, the possibility of germline variants in both genes and haploinsufficiency of *GATA2* were excluded. The occurrence of *GATA2* haploinsufficiency was ruled out in P2 by Sanger sequencing on cDNA of heterozygous coding region variants retrieved by WES. The sequencing showed that both alleles were expressed. Since P1 did not carry any heterozygous variant in the coding region, only the presence on cDNA of the coding and intronic variants reported as haploinsufficiency-causing in the Human Gene Mutation Database (HGMD, https://portal.biobase-international.com/hgmd/) could be excluded.

5.3.2.4 Screening of somatic variants in the functional domain of *STAT3*

The high-coverage targeted sequencing of exons 19, 20 and 21 of *STAT3* was performed in all six living patients (P1-P4.1, P6-P7), and four variants were identified in P1 and P2, which were patients with T-LGL. The retrieved variants in the sorted CD8+ T- and CD8- T- cell fractions (P1) or in the CD4+ T-helper cell fraction (P2) were screened. P1 carried two variants (p.(Asn647Ile), p.(Asp661Val)) in ~2% of cells in the CD8-negative fraction. The two variants were also observed in CD8+ cells with a much lower MAF. The p.Asn647Ile was present in P2 as well (CD4+ and CD8+ fractions) in addition to the p.(Ser614Arg) variant identified with a very low frequency within the CD8+ subset. The identified variants exhibited a very low frequency and were shared by several small CD8 cell clones. In addition, they were not represented in the major T-LGL clones (data not shown). This evidence led to the speculation that the identified somatic *STAT3* variants do not likely provide a major contribution to the described lymphoproliferative phenotype.

5.3.3 Characterisation of Finnish patients with the canonical DADA2 phenotype

Multiple members of a multi-case family originating from Eastern Finland (P5-P7) were diagnosed with cutaneous PAN with central nervous system complications. The described phenotypes matched the canonical clinical features originally associated with DADA2 (Zhou et *al.* 2014, Elkan et *al.* 2014). Of interest, a 40-year-long follow-up documented that all the vascular flares experienced by patients occurred in a short time span after carefully reported bacterial infections, regardless of the treatments applied. The infections mostly involved the respiratory tract or the teeth. The possible correlation between bacterial infections and the occurrence of vascular flares might be of interest for the clinical management of patients with the 'classic' vasculitic DADA2 phenotype. Clinical evaluations excluded the occurrence of lymphoproliferation in P6 and P7 up to their late 30s. P5-P7 carried two different heterozygous *ADA2* variants: the aforementioned p.Arg169Gln and a novel donor-splice variant (c.973-2A>G). The mutation screening of the transcripts of *ADA2* in the living patients (P6, P7) revealed nonsense-mediated decay (NMD) prompted by a premature stop codon after the c.973-2A>G variant. Therefore, the ensuing mono-

allelic expression of p.Arg169Gln might resemble its homozygousity observed in the previous four patients with lymphoproliferative phenotypes (P1-3, P4.2). This might suggest the possibility of the lack of one *ADA2* allele or of other genetic and/or environmental factors to modify the phenotype.

5.4 NGS-based diagnostics of genetically and clinically heterogeneous telomere biology disorders (III)

5.4.1 Analysis of NGS data revealed mutations in genes associated with telomere biology disorders (TBD)

In this study, using 'hypothesis-free' WES we have identified disease-causing variants in four patients presenting with varied immunological and/or haematological phenotypes (Table 5). In four patients were detected causal variants in genes associated with telomere biology disorders (TBDs), providing genotype-phenotype correlations for heterogeneous syndromes.

5.4.1.1 Identification of TBD-causing variants in patients with different phenotypes

All the studied patients who originated from non-consanguineous Finnish families were remitted to clinical WES to obtain a molecular diagnosis.

Patient 1 (P1.1) was a 24-year-old male presenting with chronic mucocutaneous candidiasis. The sequencing revealed a novel variant in the *DKC1* gene: c.1218_1219insCAG, p.(Asp406_Ser407insGln) (Ensembl ENSG00000130826, ENST00000369550, GRCh37.p13). The patient had been diagnosed in early childhood with vesicoureteral reflux and urinary tract infections, but no severe infections were ever documented. Since an early age, the fingernails were abnormal and easy to break, and he was subsequently diagnosed with poikiloderma (reddish skin in the neck and upper thorax). After the age of 10-13 years, recurrent aphthous ulcers on the oral

mucosa and leucoplakia (dorsal and ventral sides of the tongue) were reported. Biopsies of epithelial tissues supported a diagnosis of leucoplakia and candidiasis, requiring a further immunological workup showing only a slight decrease in IgG2 concentration and CD4+ T cells. The extension of the analysis to relatives showed similar symptoms in an older brother (P1.2), presenting with lightly milder skin, abnormalities of nails and oral mucous membranes and recurrent genital candida infections. Research of the familial *DKC1* mutation revealed P1.2 as a carrier. Due to the inferred X-linked recessive inheritance, clinical information about the maternal grandfather (died traumatically in his 40s), who was reported to have nail and skin phenotypes as well, was revisited.

Patient 2 (P2) was an 11-year-old female carrying two heterozygous variants in *TERT*: c.2051A>G, p.(Asp684Gly) (PMID:29483670) and c.3202G>A, p.Glu1068Lys

(PMID:28192371) (Ensembl ENSG00000164362, ENST00000310581, GRCh37.p13). The segregation analysis showed the two variants were compound heterozygotes, confirming the autosomal recessive segregation pattern of the phenotype. The patient was born at term, although she was small for gestational age. She experienced normal growth stages by the age of two. Developmental delay was not recorded during infancy or later. After the age of four, disorders in swallowing and oesophageal strictures were diagnosed, necessitating dilatation interventions. At the same time, mild thrombocytopenia and macrocytosis were observed in the peripheral blood; nonetheless, the bone marrow showed normal hematopoietic functions. A re-evaluation of the bone marrow was performed at the age of 10 years because cytopenia persisted, displaying features compatible with a diagnosis of myelodysplasia and hypoplastic anaemia with the exclusion of paroxysmal nocturnal haemoglobinuria and Fanconi anaemia. The extensive immunological evaluation did not detect any feature of immunodeficiency. No pulmonary, nail or skin symptoms were reported except the description of fair and thin hair. P2's older sister (16 years old) did not show any relevant symptoms, and the other close relatives were never diagnosed with cancer or had reported DKC features.

Patient 3 (P3) was a 2-year-old boy presenting with severe immunodeficiency and opportunistic infections. Two novel compound heterozygous variants in *RTEL1* (Ensembl ENSG00000258366; ENST00000318100, GRCh37.p13)) were detected (c.1721G>C, p.Arg574Pro, confirmed at RNA transcript level, and c.3724_3725delTG, p.(Cys1242Cysfs*18). The segregation analysis confirmed an autosomal recessive inheritance pattern. The patient was born prematurely and was small for the gestational age (week 31; weight -3,6 SD/ height -2,2 SD). By three months of age, he received normal vaccinations. After he was hospitalised at the age of seven months due to fever and pneumonia, *Pneumocystis jirovecii*, metapneumovirus and coronavirus were revealed by respiratory secretions. Furthermore, DNAemia was noted in high copy numbers along with retinitis. The consequent immunological work-up depicted a phenotype of $T^-B^-NK^-$ severe combined immunodeficiency (SCID). In addition, brain magnetic resonance imaging (MRI) detected microcephalia and atrophic changes in the cerebellum. Due to SCID, the patient underwent haematopoietic stem cell transplantation (HSCT) with an umbilical cord graft

at the age of nine months. The engraftment showed full donor chimerism with no manifestations of graft-versus-host disease. A post-transplant transient reactivation of cytomegalovirus was successfully managed by antiviral treatments. Before and after HSCT, the patient experienced prolonged diarrhoea and colitis-like symptoms requiring long-term total parenteral nutrition. The pulmonary problems remitted with recovering immunity without any sign of decreased liver function. After HSCT, the patient was treated by a paediatric neurologist, and genetic counselling was provided to the parents. No (known) family history of immunodeficiency or malignancy emerged. Due to the young age of the patient and the HSCT performed with chemotherapeutic conditioning, it was not possible to reliably evaluate the hair and skin features.

Patient	Sex	Age	Gene variant	Family history	Skin, nail anomalies	Hair graying/ loss	Oral mucosal changes	Aplastic anemia/ cytopenias	Immuno-deficiency	Pulmonary fibrosis	Infertility	Gastro-intestinal problems	Developmental defects	Relative telomere length*
P1.1	**M**	**24**	***DKC1***	+	+	-	+	-	+	-	**NA**	-	-	**short**
P1.2	**M**	**28**	***DKC1***	+	+	-	+	-	+	-	**NA**	-	-	**short**
P2	**F**	**11**	***TERT***	**-**	-	-	-	+	-	-	**NA**	+	+	**short**
P3	**M**	**2**	***RTEL1***	**-**	-	-	-	-	+	-	**NA**	+	+	**short**

Table 5. Demographic and clinical data of patients with germline variants in genes associated with telomere biology disorders.
Abbreviations: F, female; M, male; +, feature present; -, feature absent; NA, not available. * Compared to 113 healthy controls (age 0-83 years)

5.4.2 Functional evaluation confirming the defective telomerase function

After the identification of variants that potentially affect telomerase function, the length of the telomeres was investigated in all the affected subjects. All the patients presented shorter telomeres than the matched healthy controls, confirming the hypothesis of telomeropathy (Figure 16). In particular, the functional assessment of telomere length supported the clinical suspicion of DKC in P1.1-2, bone marrow failure in P2 and, combined with the molecular findings, the diagnosis of Hoyeraal-Hreidarsson syndrome in P3, initially diagnosed with SCID.

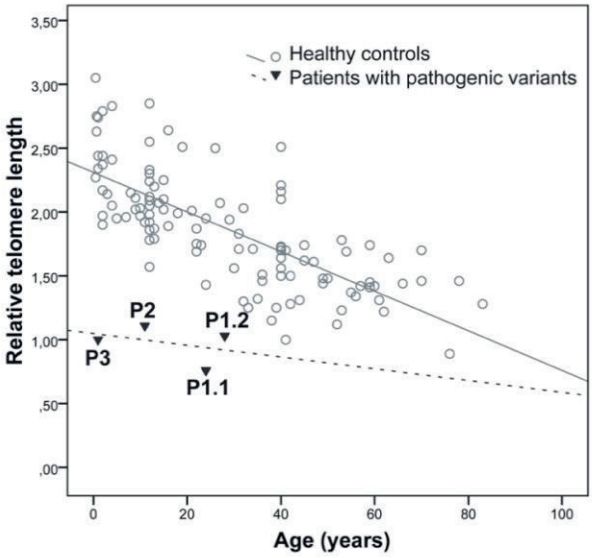

Figure 16. Relative telomere length of the patients with telomere biology disorders.
The relative telomere length (RTL) value of each sample (Y-axis) is plotted against the individual's age (X-axis). Solid triangles indicate the four TBD-patients of the study (P1.1-P3).
Open circles represent the 113 healthy controls (age 0-83 years) used for the comparison.

5.4.3 Achieving a molecular diagnosis in patients with an atypical phenotype

The detection of TBD-causing variants in patients with seemingly distinct immune or haematological features highlights the effectiveness of WES to resolve atypical traits of monogenic syndromes. Among the cohort, only two patients (P1.1, P2) displayed mucocutaneous features suggestive of DKC. Moreover, in one case (P3), the genetic analysis allowed for reaching a prompt diagnosis of HHS, although the clinical features initially described (SCID) could have

determined mistargeted approaches, impairing the possibility to identify the causal variants. Following the molecular diagnosis, the allelic segregation of the identified disease-causing variant(s) should be assessed to provide genetic counselling to the families. This is especially relevant in the case of X-linked TBDs in which the disease-causing variant may be inherited by asymptomatic female carriers.

In addition to providing a prompt molecular diagnosis, WES allowed for expanding the phenotypic spectrum of the studied diseases and for screening the secondary causative variants potentially contributing to the atypical features. From a clinical perspective, a diagnosis of telomeropathy in subjects with atypical presentations highlights the need to monitor other potential complications, such as aplastic anaemia or cytopenia. Furthermore, patients may benefit from treatment with anabolic steroids to improve cell counts instead of immunosuppressive therapy. Finally, haematopoietic stem cell transplantation can be considered for patients with severe bone marrow failure.

6 CONCLUDING REMARKS AND FUTURE PROSPECTS

The achievement of a means to diagnose rare genetic conditions is crucial for patients and families to improve genetic counselling and disease-management strategies. Identifying disease-causing alterations can enable precision medicine and can prevent unnecessary or potentially harmful interventions and treatments. Furthermore, deciphering the genetics of rare clinical phenotypes can provide insight into the underlying pathogenic mechanisms, increasing the understanding of common diseases as well. Nonetheless, not all causes of rare disorders are revealed, potentially leading to a 'diagnostic odyssey'. Hence, the aim of this study was to identify the genetic causes of rare immunological and haematological conditions overlapping PIDs in Finnish patients lacking molecular characterisation. To achieve this aim, a hypothesis-free WES approach was applied for the routine diagnostics of PIDs, and the distribution of rare causal mutations at the population level was studied.

A founder mutation in the *AICDA* gene accounting for all the cases of a rare primary antibody deficiency diagnosed in Finland thus far was identified (Hyper-IgM syndrome type 2, HIGM2). The causal allele is over-represented in the population as enclosed in an ancestral haplotype. The enrichment and the distribution of the *AICDA* founder mutation (all carriers originated from the late settlement areas) strongly reflect the genetic history of Finland and its known inhabitation patterns. Finns originated from small pools of founders, high isolation and several population bottlenecks, which caused the loss of some rare disease-causing variants and the enrichment of others. This phenomenon has been defined as FDH, which has caused some rare disorders to be more frequent in Finland than in any other country (if present). Therefore, due to the described prevalence of HIGM2 and the enrichment of the causal *AICDA v*ariants in Finland, AID-deficiency should be considered part of FDH.

In a second study, bi-allelic *ADA2* mutations causing DADA2 were identified in a set of patients with highly heterogeneous phenotypes. A multi-case family showed the classic DADA2 phenotype (cutaneous Polyarteritis nodosa [cPAN], and central nervous system complications). All the other study subjects lacked the cPAN traits but displayed varied haematological features, including cytopenia and lymphoproliferation. Notably, all the studied individuals shared an enriched-in-Finns mutation, either as homozygous or in association with a second causal variant. The patients with the DADA2-vasculitic phenotype showed only the mono-allelic expression of the variant (driven by a splicing alteration) compared with the homozygousity occurring in patients with lymphoproliferation. Therefore, it is speculated that additional genetic and/or environmental factors could contribute to the lymphoproliferative features. Overall, the results confirm the relevance of WES in discerning atypical and novel phenotypes in monogenic disorders, such as DADA2. Moreover, the emerging extension of the phenotypic spectrum associated with *ADA2* variants suggested that varied and severe haematological symptoms should be taken into account when treating DADA2 patients.

Finally, using WES, novel disease-causing variants in telomeropathy-associated genes (*DKC1, TERT* and *RTEL1* genes) were identified in three families with heterogeneous and severe immune and haematological phenotypes who lacked a conclusive diagnosis. A decreased telomere length was functionally ascertained in all cases. Notably, there was a clinical suspicion of TBD in only one family, and in another case, a different diagnosis was initially proposed. Overall, despite genetic and allelic heterogeneity, WES allowed for identifying genetic alterations in the patients, even in the absence of compelling pathognomonic signs. The achievement of a prompt diagnosis could allow for accurate genetic counselling for families, which is highly relevant due to the lack of efficient treatments applicable to some of the studied TBDs. Given the high proportion of novel variants identified, WES appears to be a more effective approach to diagnosing TBD patients than the NGS-panel targeting of known disease-causing variants or genes.

Continuous technological advancements in human genomics have largely expanded the possibilities for genetic testing, linking life science research with clinical practice. Consequently, the causes of several Mendelian diseases have been identified at a rapid pace. As shown for some of the PIDs, the achievement of a genotype–phenotype correlation can be impeded by genetic and allelic heterogeneity and by the overlap of symptoms among different rare conditions. The low prevalence can furthermore affect the clinical description of rare novel syndromes, often not encompassing the overall phenotypic spectrum of the disease, which hinders the possibility to identify other individuals with the same phenotypes. Other limitations may arise due to the high rate of novel genetic alterations, which might restrict the diagnostic target of known disease causes, as well as due to population-specific genetic structures. The resulting diagnostic flaws might result even broader because of the treatment options available to patients affected by rare diseases, which are often limited or non-existent.

By identifying disease-causing variants over highly heterogeneous clinical manifestations, WES allows for increasing the diagnostic rate of PIDs. The applied 'genotype-first' WES-based approach is an efficient method in the routine diagnostics of rare disorders, even in the absence of strong clinical suspicions. Moreover, the expansion of the disease-associated phenotypes, including atypical manifestations, represents the outset for targeted therapies for patients. Considering the relatively frequent occurrence of novel or familial disease-causing variants, the findings suggest that the application of WES (or WGS) is more effective than targeted approaches for specific phenotypes. However, despite the auspicious glimpse into the future of genetic testing applied to monogenic conditions, several limitations must be overcome to improve efficiency, principally at the level of the interpretation of clinical significance of genetic alterations which remains a bottleneck, especially in the clinical practice. The increasing availability of large-scale sequencing studies and tailored statistical approaches can support the identification of individuals' genetic variations in the frame of clinical phenotypes and at the population level. In addition, functional validation and genome editing represent the frontline in improving the understanding of the extent of genetic alterations. Genotype-phenotype correlations can be established by the evaluation of the genetic variants in model systems or using patient-derived cells, such as fibroblasts. Among the most promising technologies, novel genomic editing tools such as

CRISPR-Cas9 have a significant potential to reveal the biological impacts of rare genomic variants.

7 REFERENCES

1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature 2015 Oct 1;526(7571):68-74.

Abraham RS, Gibson LE. Error in Complementary DNA Nomenclature in "Dermatologic Features of ADA2 Deficiency in Cutaneous Polyarteritis Nodosa". JAMA Dermatol 2016 Sep 1;152(9):1065.

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods 2010 Apr;7(4):248-249.

Albers CA, Lunter G, MacArthur DG, McVean G, Ouwehand WH, Durbin R. Dindel: accurate indel calls from short-read data. Genome Res 2011 Jun;21(6):961-973.

Al-Herz W, Bousfiha A, Casanova JL, Chatila T, Conley ME, Cunningham-Rundles C, et al. Primary immunodeficiency diseases: an update on the classification from the international union of immunological societies expert committee for primary immunodeficiency. Front Immunol 2014 Apr 22;5:162.

Alsultan A, Basher E, Alqanatish J, Mohammed R, Alfadhel M. Deficiency of ADA2 mimicking autoimmune lymphoproliferative syndrome in the absence of livedo reticularis and vasculitis.

Arnheim N, Calabrese P. Understanding what determines the frequency and pattern of human germline mutations. Nat Rev Genet 2009 Jul;10(7):478-488.

Ballew BJ, Savage SA. Updates on the biology and management of dyskeratosis congenita and related telomere biology disorders. Expert Rev Hematol 2013 Jun;6(3):327-337.

Batu ED, Karadag O, Taskiran EZ, Kalyoncu U, Aksentijevich I, Alikasifoglu M, et al. A Case Series of Adenosine Deaminase 2-deficient Patients Emphasizing Treatment and Genotype-phenotype Correlations. J Rheumatol 2015 Aug;42(8):1532-1534.

Bell CJ, Dinwiddie DL, Miller NA, Hateley SL, Ganusova EE, Mudge J, et al. Carrier testing for severe childhood recessive diseases by next-generation sequencing. Sci Transl Med 2011 Jan 12;3(65):65ra4.

Belot A, Wassmer E, Twilt M, Lega JC, Zeef LA, Oojageer A, et al. Mutations in CECR1 associated with a neutrophil signature in peripheral blood. Pediatr Rheumatol Online J 2014 Sep 24;12:44-0096-12-44. eCollection 2014.

Ben-Ami T, Revel-Vilk S, Brooks R, Shaag A, Hershfield MS, Kelly SJ, et al. Extending the Clinical Phenotype of Adenosine Deaminase 2 Deficiency. J Pediatr 2016 Oct;177:316-320.

Bertuch AA. The molecular genetics of the telomere biology disorders. RNA Biol 2016 Aug 2;13(8):696-706.

Bonilla FA, Khan DA, Ballas ZK, Chinen J, Frank MM, Hsu JT, et al. Practice parameter for the diagnosis and management of primary immunodeficiency. J Allergy Clin Immunol 2015 Nov;136(5):1186-205.e1-78.

Bousfiha AA, Jeddane L, Ailal F, Benhsaien I, Mahlaoui N, Casanova JL, et al. Primary immunodeficiency diseases worldwide: more common than generally thought. J Clin Immunol 2013 Jan;33(1):1-7.

Bousfiha A, Jeddane L, Picard C, Ailal F, Bobby Gaspar H, Al-Herz W, et al. The 2017 IUIS Phenotypic Classification for Primary Immunodeficiencies. J Clin Immunol 2018 Jan;38(1):129-143.

Boycott KM, Vanstone MR, Bulman DE, MacKenzie AE. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. Nature Reviews Genetics 2013 09/03;14:681.

Bras J, Guerreiro R, Santo GC. Mutant ADA2 in vasculopathies. N Engl J Med 2014 Jul 31;371(5):478-480.

Buckley RH. Molecular defects in human severe combined immunodeficiency and approaches to immune reconstitution. Annu Rev Immunol 2004;22:625-655.

Buermans HPJ, den Dunnen JT. Next generation sequencing technology: Advances and applications. Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease; From genome to function 2014 10/01;1842(10):1932-1941.

Caorsi R, Penco F, Grossi A, Insalaco A, Omenetti A, Alessio M, et al. ADA2 deficiency (DADA2) as an unrecognised cause of early onset polyarteritis nodosa and stroke: a multicentre national study. Ann Rheum Dis 2017 Oct;76(10):1648-1656.

Cawthon RM. Telomere measurement by quantitative PCR. Nucleic Acids Res 2002 May 15;30(10):e47.

Chang X, Wang K. wANNOVAR: annotating genetic variants for personal genomes via the web. J Med Genet 2012 Jul;49(7):433-436.

Choi SK, Yoon SR, Calabrese P, Arnheim N. A germ-line-selective advantage rather than an increased mutation rate can explain some unexpectedly common human disease mutations. Proc Natl Acad Sci U S A 2008 Jul 22;105(29):10143-10148.

Chong JX, Buckingham KJ, Jhangiani SN, Boehm C, Sobreira N, Smith JD, et al. The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. Am J Hum Genet 2015 Aug 6;97(2):199-215.

Cipe FE, Aydogmus C, Serwas NK, Keskindemirci G, Boztug K. Novel Mutation in CECR1 Leads to Deficiency of ADA2 with Associated Neutropenia. J Clin Immunol 2018 Apr;38(3):273-277.

Cogne M. Activation-induced deaminase in B lymphocyte maturation and beyond. Biomed J 2013 Nov-Dec;36(6):259-268.

Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, et al. Origins and functional impact of copy number variation in the human genome. Nature 2010 Apr 1;464(7289):704-712.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 2011 May;43(5):491-498.

Dodge JA, Chigladze T, Donadieu J, Grossman Z, Ramos F, Serlicorni A, et al. The importance of rare diseases: from the gene to society. Arch Dis Child 2011 Sep;96(9):791-792.

Dokal I. Dyskeratosis congenita in all its forms. Br J Haematol 2000 Sep;110(4):768-779.

Durandy A, Cantaert T, Kracker S, Meffre E. Potential roles of activation-induced cytidine deaminase in promotion or prevention of autoimmunity in humans. Autoimmunity 2013 Mar;46(2):148-156.

Durandy A, Taubenheim N, Peron S, Fischer A. Pathophysiology of B-cell intrinsic immunoglobulin class switch recombination deficiencies. Adv Immunol 2007;94:275-306.

Elbracht M, Mull M, Wagner N, Kuhl C, Abicht A, Kurth I, et al. Stroke as Initial Manifestation of Adenosine Deaminase 2 Deficiency. Neuropediatrics 2017 Apr;48(2):111-114.

ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 2012 Sep 6;489(7414):57-74.

Ezkurdia I, Juan D, Rodriguez JM, Frankish A, Diekhans M, Harrow J, et al. Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes. Hum Mol Genet 2014 Nov 15;23(22):5866-5878.

Fellmann F, Angelini F, Wassenberg J, Perreau M, Arenas Ramirez N, Simon G, et al. IL-17 receptor A and adenosine deaminase 2 deficiency in siblings with recurrent infections and chronic inflammation. J Allergy Clin Immunol 2016 Apr;137(4):1189-1196.e2.

Fietta P. Autoinflammatory diseases: the hereditary periodic fever syndromes. Acta Biomed 2004 Aug;75(2):92-99.

Franco R, Pacheco R, Gatell JM, Gallart T, Lluis C. Enzymatic and extraenzymatic role of adenosine deaminase 1 in T-cell-dendritic cell contacts and in alterations of the immune function. Crit Rev Immunol 2007;27(6):495-509.

Garg N, Kasapcopur O, Foster J,2nd, Barut K, Tekin A, Kizilkilic O, et al. Novel adenosine deaminase 2 mutations in a child with a fatal vasculopathy. Eur J Pediatr 2014 Jun;173(6):827-830.

Glousker G, Touzot F, Revy P, Tzfati Y, Savage SA. Unraveling the pathogenesis of Hoyeraal-Hreidarsson syndrome, a complex telomere biology disorder. Br J Haematol 2015 Aug;170(4):457-471.

Gonzalez Santiago TM, Zavialov A, Saarela J, Seppanen M, Reed AM, Abraham RS, et al. Dermatologic Features of ADA2 Deficiency in Cutaneous Polyarteritis Nodosa. JAMA Dermatol 2015 Nov;151(11):1230-1234.

Goriely A, McVean GA, Rojmyr M, Ingemarsson B, Wilkie AO. Evidence for selective advantage of pathogenic FGFR2 mutations in the male germ line. Science 2003 Aug 1;301(5633):643-646.

Gulino AV, Notarangelo LD. Hyper IgM syndromes. Curr Opin Rheumatol 2003 Jul;15(4):422-429.

Günthner R, Wagner M, Thurm T, Ponsel S, HÃ¶fele J, Lange-Sperandio B. Identification of co-occurrence in a patient with Dent's disease and ADA2-deficiency by exome sequencing. Gene 2018 04/05;649:23-26.

Haffner ME, Whitley J, Moses M. Two decades of orphan product development. Nat Rev Drug Discov 2002 Oct;1(10):821-825.

Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 2005 Jan 1;33(Database issue):D514-7.

Hashem H, Egler R, Dalal J. Refractory Pure Red Cell Aplasia Manifesting as Deficiency of Adenosine Deaminase 2. J Pediatr Hematol Oncol 2017 Jul;39(5):e293-e296.

Hashem H, Kelly SJ, Ganson NJ, Hershfield MS. Deficiency of Adenosine Deaminase 2 (DADA2), an Inherited Cause of Polyarteritis Nodosa and a Mimic of Other Systemic Rheumatologic Disorders. Curr Rheumatol Rep 2017 Oct 5;19(11):70-017-0699-8.

Hashem H, Kumar AR, Muller I, Babor F, Bredius R, Dalal J, et al. Hematopoietic stem cell transplantation rescues the hematological, immunological, and vascular phenotype in DADA2. Blood 2017 Dec 14;130(24):2682-2688.

Hastings IM. Germline selection: population genetic aspects of the sexual/asexual life cycle. Genetics 1991 Dec;129(4):1167-1176.

Heiss NS, Knight SW, Vulliamy TJ, Klauck SM, Wiemann S, Mason PJ, et al. X-linked dyskeratosis congenita is caused by mutations in a highly conserved gene with putative nucleolar functions. Nat Genet 1998 May;19(1):32-38.

Hsu AP, West RR, Calvo KR, Cuellar-Rodriguez J, Parta M, Kelly SJ, et al. Adenosine deaminase type 2 deficiency masquerading as GATA2 deficiency: Successful hematopoietic stem cell transplantation. J Allergy Clin Immunol 2016 Aug;138(2):628-630.e2.

International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature 2004 Oct 21;431(7011):931-945.

Jain A, Misra DP, Sharma A, Wakhlu A, Agarwal V, Negi VS. Vasculitis and vasculitis-like manifestations in monogenic autoinflammatory syndromes. Rheumatol Int 2017 Oct 14.

Jamuar SS, Tan EC. Clinical application of next-generation sequencing for Mendelian diseases. Hum Genomics 2015 Jun 16;9:10-015-0031-5.

Kaariainen H, Muilu J, Perola M, Kristiansson K. Genetics in an isolated population like Finland: a different basis for genomic medicine? J Community Genet 2017 Oct;8(4):319-326.

Kaljas Y, Liu C, Skaldin M, Wu C, Zhou Q, Lu Y, et al. Human adenosine deaminases ADA1 and ADA2 bind to different subsets of immune cells. Cellular and Molecular Life Sciences 2017 02/01;74(3):555-570.

Keer N, Hershfield M, Caskey T, Unizony S. Novel compound heterozygous variants in CECR1 gene associated with childhood onset polyarteritis nodosa and deficiency of ADA2. Rheumatology (Oxford) 2016 Jun;55(6):1145-1147.

Keim C, Kazadi D, Rothschild G, Basu U. Regulation of AID, the B-cell genome mutator. Genes Dev 2013 Jan 1;27(1):1-17.

Kim BY, Park JH, Jo HY, Koo SK, Park MH. Optimized detection of insertions/deletions (INDELs) in whole-exome sequencing data. PLoS One 2017 Aug 9;12(8):e0182272.

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 2014 Mar;46(3):310-315.

Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, et al. Rate of de novo mutations and the importance of father's age to disease risk. Nature 2012 Aug 23;488(7412):471-475.

Krawczak M, Cooper DN. The human gene mutation database. Trends Genet 1997 Mar;13(3):121-122.

Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 2009;4(7):1073-1081.

Kumar R, DiMenna L, Schrode N, Liu TC, Franck P, Munoz-Descalzo S, et al. AID stabilizes stem-cell phenotype by removing epigenetic memory of pluripotency genes. Nature 2013 Aug 1;500(7460):89-92.

Kuraoka M, Holl TM, Liao D, Womble M, Cain DW, Reynolds AE, et al. Activation-induced cytidine deaminase mediates central tolerance in B cells. Proc Natl Acad Sci U S A 2011 Jul 12;108(28):11560-11565.

Lamprecht P, Humrich JY, Diebold I, Riemekasten G. Diagnosis of deficiency of adenosine deaminase 2 with early onset polyarteritis nodosa in an adult patient with a novel compound heterozygous CECR1 mutation. Clin Exp Rheumatol 2018 Mar 15.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature 2001 Feb 15;409(6822):860-921.

Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature 2016 Aug 18;536(7616):285-291.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009 Aug 15;25(16):2078-2079.

Lim BC, Yoo SK, Lee S, Shin JY, Hwang H, Chae JH, et al. Hoyeraal-Hreidarsson syndrome with a DKC1 mutation identified by whole-exome sequencing. Gene 2014 Aug 10;546(2):425-429.

Lim ET, Wurtz P, Havulinna AS, Palta P, Tukiainen T, Rehnstrom K, et al. Distribution and medical impact of loss-of-function variants in the Finnish founder population. PLoS Genet 2014 Jul 31;10(7):e1004494.

Lionel AC, Costain G, Monfared N, Walker S, Reuter MS, Hosseini SM, et al. Improved diagnostic yield compared with targeted gene sequencing panels suggests a role for whole-genome sequencing as a first-tier genetic test. Genet Med 2017 Aug 3.

Lohmueller KE, Indap AR, Schmidt S, Boyko AR, Hernandez RD, Hubisz MJ, et al. Proportionally more deleterious genetic variation in European than in African populations. Nature 2008 Feb 21;451(7181):994-997.

Makitie O, Perheentupa J, Kaitila I. Growth in cartilage-hair hypoplasia. Pediatr Res 1992 Feb;31(2):176-180.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010 Sep;20(9):1297-1303.

McKusick VA. Mendelian Inheritance in Man and its online version, OMIM. Am J Hum Genet 2007 Apr;80(4):588-604.

Meyers G, Ng YS, Bannock JM, Lavoie A, Walter JE, Notarangelo LD, et al. Activation-induced cytidine deaminase (AID) is required for B-cell tolerance in humans. Proc Natl Acad Sci U S A 2011 Jul 12;108(28):11554-11559.

Michniacki TF, Hannibal M, Ross CW, Frame DG, DuVall AS, Khoriaty R, et al. Hematologic Manifestations of Deficiency of Adenosine Deaminase 2 (DADA2) and Response to Tumor Necrosis Factor Inhibition in DADA2-Associated Bone Marrow Failure. J Clin Immunol 2018 02/06.

Minegishi Y, Lavoie A, Cunningham-Rundles C, Bedard PM, Hebert J, Cote L, et al. Mutations in activation-induced cytidine deaminase in patients with hyper IgM syndrome. Clin Immunol 2000 Dec;97(3):203-210.

Nachman MW, Crowell SL. Estimate of the mutation rate per nucleotide in humans. Genetics 2000 Sep;156(1):297-304.

Nanthapisal S, Murphy C, Omoyinmi E, Hong Y, Standing A, Berg S, et al. Deficiency of Adenosine Deaminase Type 2: A Description of Phenotype and Genotype in Fifteen Cases. Arthritis Rheumatol 2016 Sep;68(9):2314-2322.

Navon Elkan P, Pierce SB, Segel R, Walsh T, Barash J, Padeh S, et al. Mutant Adenosine Deaminase 2 in a Polyarteritis Nodosa Vasculopathy. N Engl J Med 2014 03/06; 2018/05;370(10):921-931.

Neufeld M, Maclaren NK, Blizzard RM. Two types of autoimmune Addison's disease associated with different polyglandular autoimmune (PGA) syndromes. Medicine (Baltimore) 1981 Sep;60(5):355-362.

Norberg A, Rosen A, Raaschou-Jensen K, Kjeldsen L, Moilanen JS, Paulsson-Karlsson Y, et al. Novel variants in Nordic patients referred for genetic testing of telomere-related disorders. Eur J Hum Genet 2018 Jun;26(6):858-867.

Norio R. Finnish Disease Heritage I: characteristics, causes, background. Hum Genet 2003 May;112(5-6):441-456.

Norio R. The Finnish Disease Heritage III: the individual diseases. Hum Genet 2003 May;112(5-6):470-526.

Oishi M, Oishi A, Gotoh N, Ogino K, Higasa K, Iida K, et al. Comprehensive molecular diagnosis of a large cohort of Japanese retinitis pigmentosa and Usher syndrome patients by next-generation sequencing. Invest Ophthalmol Vis Sci 2014 Oct 16;55(11):7369-7375.

Omoyinmi E, Standing A, Keylock A, Price-Kuehne F, Melo Gomes S, Rowczenio D, et al. Clinical impact of a targeted next-generation sequencing gene panel for autoinflammation and vasculitis. PLoS One 2017 Jul 27;12(7):e0181874.

Ozkurede VU, Franchi L. Immunology in clinic review series; focus on autoinflammatory diseases: role of inflammasomes in autoinflammatory syndromes. Clin Exp Immunol 2012 Mar;167(3):382-390.

Pabinger S, Dander A, Fischer M, Snajder R, Sperk M, Efremova M, et al. A survey of tools for variant analysis of next-generation genome sequencing data. Brief Bioinform 2014 Mar;15(2):256-278.

Peltonen L, Jalanko A, Varilo T. Molecular genetics of the Finnish disease heritage. Hum Mol Genet 1999;8(10):1913-1923.

Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. Genic intolerance to functional variation and the interpretation of personal genomes. PLoS Genet 2013;9(8):e1003709.

Picard C, Al-Herz W, Bousfiha A, Casanova JL, Chatila T, Conley ME, et al. Primary Immunodeficiency Diseases: an Update on the Classification from the International Union of Immunological Societies Expert Committee for Primary Immunodeficiency 2015. J Clin Immunol 2015 Nov;35(8):696-726.

Picard C, Bobby Gaspar H, Al-Herz W, Bousfiha A, Casanova JL, Chatila T, et al. International Union of Immunological Societies: 2017 Primary Immunodeficiency Diseases Committee Report on Inborn Errors of Immunity. J Clin Immunol 2018 Jan;38(1):96-128.

Pogue RE, Cavalcanti DP, Shanker S, Andrade RV, Aguiar LR, de Carvalho JL, et al. Rare genetic diseases: update on diagnosis, treatment and online resources. Drug Discovery Today 2017 Available online 10 November 2017.

Polvi A, Linturi H, Varilo T, Anttonen AK, Byrne M, Fokkema IF, et al. The Finnish disease heritage database (FinDis) update-a database for the genes mutated in the Finnish disease heritage brought to the next-generation sequencing era. Hum Mutat 2013 Nov;34(11):1458-1466.

Poswar Fde O, da Fonseca RM, de Albuquerque LC, Zhou Q, Jardim LB, Monte TL, et al. Adenosine deaminase 2 deficiency presenting as spastic paraplegia and systemic vasculitis. J Neurol 2016 Apr;263(4):818-820.

Qin J, Calabrese P, Tiemann-Boege I, Shinde DN, Yoon SR, Gelfand D, et al. The molecular anatomy of spontaneous germline mutations in human testes. PLoS Biol 2007 Sep;5(9):e224.

Quartier P, Bustamante J, Sanal O, Plebani A, Debre M, Deville A, et al. Clinical, immunologic and genetic analysis of 29 patients with autosomal recessive hyper-IgM syndrome due to Activation-Induced Cytidine Deaminase deficiency. Clin Immunol 2004 Jan;110(1):22-29.

Rabbani B, Tekin M, Mahdieh N. The promise of whole-exome sequencing in medical genetics. J Hum Genet 2014 Jan;59(1):5-15.

Revy P, Geissmann F, Debre M, Fischer A, Durandy A. Normal CD40-mediated activation of monocytes and dendritic cells from patients with hyper-IgM syndrome due to a CD40 pathway defect in B cells. Eur J Immunol 1998 Nov;28(11):3648-3654.

Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med 2015 May;17(5):405-424.

Ridanpaa M, van Eenennaam H, Pelin K, Chadwick R, Johnson C, Yuan B, et al. Mutations in the RNA component of RNase MRP cause a pleiotropic human disease, cartilage-hair hypoplasia. Cell 2001 Jan 26;104(2):195-203.

Rogozin IB, Pavlov YI. Theoretical analysis of mutation hotspots and their DNA sequence context specificity. Mutat Res 2003 Sep;544(1):65-85.

Sahin S, Adrovic A, Barut K, Ugurlu S, Turanli ET, Ozdogan H, et al. Clinical, imaging and genotypical features of three deceased and five surviving cases with ADA2 deficiency. Rheumatol Int 2018 01/01;38(1):129-136.

Sandin S, Rhodes D. Telomerase structure. Current Opinion in Structural Biology 2014 April 2014;25:104-110.

Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol 1975 May 25;94(3):441-448.

Schepp J, Bulashevska A, Mannhardt-Laakmann W, Cao H, Yang F, Seidl M, et al. Deficiency of Adenosine Deaminase 2 Causes Antibody Deficiency. J Clin Immunol 2016 Apr;36(3):179-186.

Schepp J, Proietti M, Frede N, Buchta M, Hubscher K, Rojas Restrepo J, et al. Screening of 181 Patients With Antibody Deficiency for Deficiency of Adenosine Deaminase 2 Sheds New Light on the Disease in Adulthood. Arthritis Rheumatol 2017 Aug;69(8):1689-1700.

Schwarz JM, Cooper DN, Schuelke M, Seelow D. MutationTaster2: mutation prediction for the deep-sequencing age. Nat Methods 2014 Apr;11(4):361-362.

Seifi M, Walter MA. Accurate prediction of functional, structural, and stability changes in PITX2 mutations using in silico bioinformatics algorithms. PLoS One 2018 Apr 17;13(4):e0195971.

Seleman M, Hoyos-Bachiloglu R, Geha RS, Chou J. Uses of Next-Generation Sequencing Technologies for the Diagnosis of Primary Immunodeficiencies. Front Immunol 2017 Jul 24;8:847.

Selenius JS, Martelius T, Pikkarainen S, Siitonen S, Mattila E, Pietikainen R, et al. Unexpectedly High Prevalence of Common Variable Immunodeficiency in Finland. Front Immunol 2017 Sep 28;8:1190.

Shen T, Lee A, Shen C, Lin CJ. The long tail and rare disease research: the impact of next-generation sequencing for rare Mendelian disorders. Genet Res (Camb) 2015 Sep 14;97:e15.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 2001 Jan 1;29(1):308-311.

Simons YB, Turchin MC, Pritchard JK, Sella G. The deleterious mutation load is insensitive to recent population history. Nat Genet 2014 Mar;46(3):220-224.

Skrabl-Baumgartner A, Plecko B, Schmidt WM, Konig N, Hershfield M, Gruber-Sedlmayr U, et al. Autoimmune phenotype with type I interferon signature in two brothers with ADA2 deficiency carrying a novel CECR1 mutation. Pediatr Rheumatol Online J 2017 Aug 22;15(1):67-017-0193-x.

Sulonen AM, Ellonen P, Almusa H, Lepisto M, Eldfors S, Hannula S, et al. Comparison of solution-based exome capture methods for next generation sequencing. Genome Biol 2011 Sep 28;12(9):R94-2011-12-9-r94.

Tiemann-Boege I, Navidi W, Grewal R, Cohn D, Eskenazi B, Wyrobek AJ, et al. The observed human sperm mutation frequency cannot explain the achondroplasia paternal age effect. Proc Natl Acad Sci U S A 2002 Nov 12;99(23):14952-14957.

Townsley DM, Dumitriu B, Young NS. Bone marrow failure and the telomeropathies. Blood 2014 Oct 30;124(18):2775-2783.

Turner TN, Coe BP, Dickel DE, Hoekzema K, Nelson BJ, Zody MC, et al. Genomic Patterns of De Novo Mutation in Simplex Autism. Cell 2017 Oct 19;171(3):710-722.e12.

Uettwiller F, Sarrabay G, Rodero MP, Rice GI, Lagrue E, Marot Y, et al. ADA2 deficiency: case report of a new phenotype and novel mutation in two sisters. RMD Open 2016 May 16;2(1):e000236-2015-000236. eCollection 2016.

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics 2013;43:11.10.1-33.

van Driel R, Fransz PF, Verschure PJ. The eukaryotic genome: a system regulated at different hierarchical levels. J Cell Sci 2003 Oct 15;116(Pt 20):4067-4075.

Van Eyck L, Liston A, Meyts I. Mutant ADA2 in vasculopathies. N Engl J Med 2014 Jul 31;371(5):478-479.

Van Eyck L, Liston A, Wouters C. Mutant ADA2 in vasculopathies. N Engl J Med 2014 Jul 31;371(5):480.

Van Eyck L,Jr, Hershfield MS, Pombal D, Kelly SJ, Ganson NJ, Moens L, et al. Hematopoietic stem cell transplantation rescues the immunologic phenotype and prevents vasculopathy in patients with adenosine deaminase 2 deficiency. J Allergy Clin Immunol 2015 Jan;135(1):283-7.e5.

van Montfrans J, Zavialov A, Zhou Q. Mutant ADA2 in vasculopathies. N Engl J Med 2014 Jul 31;371(5):478.

Van Montfrans JM, Hartman EA, Braun KP, Hennekam EA, Hak EA, Nederkoorn PJ, et al. Phenotypic variability in patients with ADA2 deficiency due to identical homozygous R169Q mutations. Rheumatology (Oxford) 2016 May;55(5):902-910.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The Sequence of the Human Genome. Science 2001 American Association for the Advancement of Science;291(5507):1304-1351.

von der Lippe C, Diesen PS, Feragen KB. Living with a rare disorder: a systematic review of the qualitative literature. Molecular Genetics & Genomic Medicine 2017;5(6):758-773.

Vulliamy TJ, Marrone A, Knight SW, Walne A, Mason PJ, Dokal I. Mutations in dyskeratosis congenita: their impact on telomere length and the diversity of clinical presentation. Blood 2006 Apr 1;107(7):2680-2685.

Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 2010 Sep;38(16):e164.

Westendorp WF, Nederkoorn PJ, Aksentijevich I, Hak AE, Lichtenbelt KD, Braun KP. Unexplained early-onset lacunar stroke and inflammatory skin lesions: Consider ADA2 deficiency. Neurology 2015 May 19;84(20):2092-2093.

Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, et al. The complete genome of an individual by massively parallel DNA sequencing. Nature 2008 Apr 17;452(7189):872-876.

Young DL, Fields S. The role of functional data in interpreting the effects of genetic variation. Mol Biol Cell 2015 Nov 5;26(22):3904-3908.

Zavialov AV, Gracia E, Glaichenhaus N, Franco R, Zavialov AV, Lauvau G. Human adenosine deaminase 2 induces differentiation of monocytes into macrophages and stimulates proliferation of T helper cells and macrophages. J Leukoc Biol 2010 Aug;88(2):279-290.

Zeft AS, Spalding SJ. Autoinflammatory syndromes: fever is not always a sign of infection. Cleve Clin J Med 2012 Aug;79(8):569-581.

Zhou Q, Yang D, Ombrello AK, Zavialov AV, Toro C, Zavialov AV, et al. Early-onset stroke and vasculopathy associated with mutations in ADA2. N Engl J Med 2014 Mar 6;370(10):911-920.

8 ORIGINAL PUBLICATIONS

Recent Publications in this Series

84/2018 Omar Youssef
Analysis of Gene Mutations in Exhaled Breath Condensate from Healthy and Lung Cancer Individuals and Profiling of Mutations and Gut Microbiota in Stools from Patients with Gastrointestinal Neoplasms
85/2018 Teija Kasteenpohja
Treatment Seeking, Treatment Adequacy and Outcome of Depressive and Anxiety Disorders among Young Adults in Finland — Findings from a Population-Based Sample
86/2018 Farhana Jahan
Phosphorylation of the α- and β-Chains of LFA-1 Regulates its Interaction with Cytoplasmic Proteins and Crosstalk to VLA-4 Integrin
87/2018 Jenny Högström-Stakem
Analysis of Molecular Pathways in Intestinal Stem Cells and Colorectal Cancer Progression
88/2018 Sanna Pallaskorpi
Long-Term Outcome of Bipolar I and II Disorders
89/2018 Laura Hintikka
Development of Mass Spectrometric Methods for Analysis of Anabolic Androgenic Steroids
90/2018 Mari Muurinen
Silver-Russell Syndrome and Human Growth: Genetic and Epigenetic Studies
91/2018 Ali Oghabian
Bioinformatics Analysis of Intron Retention Events Associated with the Minor Spliceosome
92/2018 Jenni Puurunen
Metabolomic Characterization of Canine Behavioural Disorders: Fearfulness and Hyperactivity/Impulsivity
93/2018 Marit Ilves
Immunomodulatory Effects of Engineered Nanomaterials in Healthy and Diseased Lungs and Skin
94/2018 Laura Lahdentausta
Serum and Saliva Biomarkers in Cardiovascular Diseases and Periodontitis—Smoking as a Confounding Factor
95/2018 Joni Lindbohm
Unfolding Lipid Profile- and Sex-Paradoxes in Epidemiology of Subarachnoid Haemorrhage
96/2018 Anna But
Mitigating Bias and Dealing with Multiple Time Scales in Cohort Studies — Studying Medications and Complications of Diabetes
97/2018 Ashwini S. Nagaraj
Unravelling Lung Cancer Heterogeneity and Associated Therapeutic Responses Using *In Vivo* and *Ex Vivo* Model Systems
98/2018 Lotta Schepel
Strategies for Medication Safety: An Organization-Based Approach Focusing on High-Alert Medications and Clinical Pharmacy Services in Helsinki University Hospital
1/2019 Cristina Fevola
Distribution and Clinical Associations of Ljungan Virus (Parechovirus B)
2/2019 Anita Valkama
Dietary Change, Obesity, and Metabolic Markers in Pregnancy — Studies in Women at Risk for Gestational Diabetes Mellitus
3/2019 Feng Deng
Structure-Activity Relationships of Efflux Transporter Inhibitors
4/2019 Leena Yadav
Human Protein Phosphatase Interactions and Dynamics: Proteomic and Functional Perspective

Helsinki 2019 ISSN 2342-3161 ISBN 978-951-51-4783-7